HafenCity Universität Hamburg

Satellite gravimetry for climate model evaluation

Dissertation

zur Erlangung des akademischen Grades
Doktor-Ingenieur (Dr.-Ing.)

an der

HafenCity Universität Hamburg

vorgelegt von

Laura Jensen

Hamburg, Juni 2021

Gutachter:
Prof. Dr.-Ing. Annette Eicker (HafenCity Universität Hamburg)
Prof. Dr.techn. Mag.rer.nat. Roland Pail (Technische Universität München)

License notice

Acknowledgements

This thesis would not exist without the support from my colleagues, friends, and family, whom I want to thank.

My first—and by far biggest—thanks go to Annette Eicker for giving me the opportunity to carry out this dissertation project. I am deeply grateful for her dedicated supervision and her extraordinary engagement in all major and minor scientific and non-scientific issues during the whole time of my PhD. Thank you for numerous inspiring conversations and your extensive guidance!

I also thank Roland Pail for his helpful support, great collaboration, and for acting as external reviewer.

I want to especially thank Henryk Dobslaw, not only for his generous help during the process of writing the three publications, but also for his very supportive guidance and advice throughout my whole PhD. I always appreciated your clear view on things!

Furthermore, I owe many thanks to Tobias Stacke and Vincent Humphrey. The fruitful discussions with them always inspired me and brought me forward with my research. Working together with you was a pleasure for me!

I also thank my (former) colleagues at HCU, namely Felix Tschirschwitz, Simon Deggim, Thomas Kersten, Inga Schlegel, Hannes Kröger, Juiwen Chang, Kuei-Hua Hsu, and Daniel Blank for smaller or larger discussions, for helping me to solve one or another technical or scientific problem, and for always finding ways to further motivate me. You were a great company!

My wholehearted thanks go to Lester Lembke-Jene for his steady reassurance and affectionate advice whenever I encountered difficulties. Your faith in me always kept me grounded.

Big hugs and many thanks also go to my dear friends Maren, Benjo, Ribana, and Sarah for their enduring support and continuous encouragement.

I also want to thank all members of my beloved recorder ensemble Flauto Vivo. Making music together, having cake and nerdy off-topic discussions, were a constant and very welcome source of fresh energy for my work.

Finally, I am most grateful to my parents Ulrike and Jens-Peter and my brother Philipp for their true interest in my work, their unconditional support and reliance, and for just always being there. Thank you!

Abstract

Global coupled Earth System Models (ESMs) are important for predicting future climate conditions. To assess the quality and reliability of ESMs it is crucial to evaluate them against independent observations. In this thesis, land water storage-related variables from ESMs taking part in the Coupled Model Intercomparison Project Phases 5 and 6 (CMIP5 and CMIP6) are compared to observed terrestrial water storage (TWS) changes from the Gravity Recovery And Climate Experiment (GRACE) and its Follow-On (GRACE-FO) satellite missions. This thesis is the first study to provide a comprehensive assessment of ways in which space gravimetric measurements can be utilized to evaluate coupled ESMs, thereby tackling several challenges of such a comparison:

Apart from external forcing data (e.g., the Sun's energy and greenhouse gas concentrations), CMIP ESMs are not constrained by observations, but evolve freely over time after being initialized. As a consequence, over long time spans they reproduce the climate variability in a statistical sense only, but not the exact timing of particular events. Thus, when analyzing (long-term) climate projections, a direct comparison to observed time series is not feasible, but higher order metrics as the linear trend, seasonal cycle, and interannual variability have to be utilized. Furthermore, discrepancies between observed TWS (full integrated water column, possibly superimposed by non-hydrological mass changes) and modeled TWS (representing only soil moisture and snow) regionally hamper the interpretation of results.

In the comparison of a CMIP5 ESM ensemble with GRACE observations, simulated long-term wetting and drying trends are found to be consistent with observed TWS trends in several regions of the world (e.g., the Mediterranean, Southwestern United States, Central Asia). However, it is shown that interannual variations obscuring long-term trends in TWS can have a large influence over 30 years and more, which regionally prevents reliable conclusions about long-term wetting or drying from the short GRACE time series. In addition to long-term trends, the seasonal cycle and interannual variations of TWS in a CMIP6 multi-model ensemble are assessed with respect to present-day observed conditions. The model data are also analyzed for potential future changes of TWS variability, as these might be an important target for a future gravity mission with enhanced sensitivity. In contrast to long-term climate projections, decadal predictions can directly be compared on time series level, because they are frequently initialized with observed states of atmosphere and oceans. To this end, a GRACE-based TWS reconstruction is used to evaluate decadal climate predictions from CMIP5 and CMIP6 with the result that they provide skillful forecasts of TWS anomalies in markedly humid climates for two years and more into the future.

Overall, this thesis highlights mutual benefits of climate modeling and geodesy: On the one hand, satellite-observed TWS has a great potential to validate the performance or to hint at shortcomings of ESMs for land water storage-related variables. On the other hand, ESM output can provide information on expected climate signals in TWS, which is important for future plans in space gravimetry aiming at the long-term monitoring of climate variations.

Zusammenfassung

Globale gekoppelte Erdsystemmodelle (ESMs) werden für die Vorhersage zukünftiger Klimabedingungen benötigt. Eine Methode zur Beurteilung ihrer Qualität und Zuverlässigkeit ist ihre Evaluierung anhand unabhängiger Beobachtungen. In dieser Arbeit wird modellierter terrestrischer Wasserspeicher (TWS) aus ESMs, die am Coupled Model Intercomparison Project der Phasen 5 und 6 (CMIP5 und CMIP6) teilnehmen, mit beobachteten TWS-Änderungen aus Daten der Satellitenmissionen Gravity Recovery And Climate Experiment (GRACE) und GRACE Follow-On (GRACE-FO) verglichen. Dies ist die erste umfassende Studie, die Möglichkeiten des Einsatzes von Satellitengravimetrie zur Evaluierung gekoppelter ESMs untersucht. Dabei werden verschiedene Herausforderungen angegangen:

Abgesehen von externen Antriebsdaten (z. B. Sonneneinstrahlung und Treibhausgas-Konzentrationen) gehen keine Beobachtungen in CMIP-ESMs ein. Nach ihrer Initialisierung laufen sie frei über den Vorhersagezeitraum. Infolgedessen repräsentieren sie die Klimavariabilität über lange Zeiträume nur im statistischen Sinne, und können keine exakten Zeitpunkte für bestimmte Ereignisse vorhersagen. Daher ist der direkte Vergleich von (langfristigen) Klimaprojektionen mit beobachteten Zeitreihen nicht möglich, sondern es müssen Metriken höherer Ordnung, wie der lineare Trend, der Jahresgang und die interannuelle Variabilität, verwendet werden. Außerdem erschweren Diskrepanzen zwischen beobachtetem TWS (integrale Wassersäule, möglicherweise überlagert von nicht-hydrologischen Signalen) und modelliertem TWS (nur Bodenfeuchte und Schnee) die Interpretation der Ergebnisse.

Der Vergleich eines CMIP5-ESM-Ensembles mit GRACE-Beobachtungen ergibt, dass die simulierten langfristigen Trends mit den beobachteten TWS-Trends in einigen Gebieten übereinstimmen, wie z. B. im Mittelmeerraum, Südwesten der USA und Zentralasien. Jedoch zeigt sich, dass interannuelle Variationen, welche die Trends überlagern, über 30 Jahre und länger einen großen Einfluss haben können. Daher sind verlässliche Aussagen über langfristige Trends mit der kurzen GRACE-Zeitreihe in vielen Regionen derzeit noch nicht möglich. In der Arbeit werden auch der Jahresgang und interannuelle Variationen von TWS in einem CMIP6-Multi-Modell-Ensemble im Vergleich zu beobachtetem TWS untersucht. Aus den Modelldaten werden mögliche zukünftige Änderungen der TWS-Variabilität abgeleitet. Die Erfassung solcher Änderungen könnte eine wichtige Zielvorgabe für zukünftige Satellitengravimetrie-Missionen mit höherer Sensitivität darstellen. Im Gegensatz zu langfristigen Klimaprojektionen können dekadische Prädiktionen direkt auf Zeitreihenebene verglichen werden, da sie z.B. jährlich mit beobachteten Zuständen von Atmosphäre und Ozean initialisiert werden. In der Arbeit wird eine GRACE-basierte TWS-Rekonstruktion zur Evaluierung von dekadischen CMIP5- und CMIP6-Prädiktionen verwendet. Die Ergebnisse zeigen, dass die Modelle fähig sind, Vorhersagen über TWS-Anomalien (vor allem in feuchten Klimazonen) von zwei Jahren und länger in die Zukunft zu treffen.

Diese Arbeit hebt den Nutzen hervor, den Klimamodellierung und Geodäsie voneinander ziehen können: Einerseits hat satellitenbeobachteter TWS viel Potential, Qualitäten und Unzulänglichkeiten von ESMs in Bezug auf ihre Repräsentation von TWS aufzuzeigen. Andererseits können ESMs Informationen über Klimasignale in TWS liefern, was wichtig für die Planung zukünftiger Satellitengravimetrie-Missionen mit dem Ziel langfristiger Klimaüberwachung ist.

Contents

1. Introduction

1.1. Motivation and background

"The earth is a fine place and worth fighting for."
Ernest Hemingway

"Climate change is the greatest threat to our existence in our short history on this planet. Nobody's going to buy their way out of its effects."
Mark Ruffalo, Actor & Environmentalist

A profound scientific understanding of the Earth system is the fundamental basis for developing political mitigation and adaptation strategies to climate change. Science can contribute to facing climate change by collecting and interpreting observational data and by making projections about the expected future evolution of our planet. A tool especially valuable to forecast future climate conditions is the operation of climate models. Using the power of supercomputers, climate models are run as complex software programs to simulate possible future states of our environment. Results of climate models build, e.g., the basis of the Assessment Reports of the Intergovernmental Panel on Climate Change (IPCC), which are an important source of information for politicians and stakeholders making climate-related decisions. Therefore, the quality of climate models is of crucial importance to induce robust conclusions. A strategy to assess the quality of climate models and to further improve them is their evaluation by means of observations. By running climate models for the past and comparing the output to observational data, model skills and shortcomings can be revealed.

The evaluation of climate models by means of observations poses many challenges. Methods to face these challenges have to be adapted to (1) the climate variables considered, (2) the spatial, and (3) the temporal scale of the evaluation. The first issue arises due to the fact that climate variables provided by climate models, due to their simplified representation of reality, do not always describe the exact same entity as can be observed. Furthermore, observations as well as model results are affected by various sources of uncertainty that have to be taken into account. The second issue, the spatial scale, includes the resolution of climate models, which often is too coarse to resolve small-scale processes affecting climate, thereby hampering comparison to observations. Also, the spatial sampling of the observational record is crucial for meaningful evaluation. Many geophysical, geological and hydrological observations, which are often collected as in-situ point measurements, lack uniform spatial coverage. Moreover, to differentiate the quality of climate models, evaluation can be performed on a global or on a regional scale, with likely different conclusions. The third issue, the temporal scale, depends on the time horizon considered: the evaluation approaches differ if the forecast covers only a few years or rather several decades to centuries. Furthermore, the length of the observational time series is an important factor for model evaluation. As climate is defined as the long-term average of weather, with a typical averaging period of 30 years (according to the World Meteorological Organization, WMO), this would in principle be the minimum length for a data record to study climate change. However, even if a time series of observations is still too short to carry out dedicated climate studies, it can already be used for constraining variables in climate models. Especially remotely sensed satellite observations, which often have the advantage of a global coverage, are in most cases only available for time spans of less than 30 years, but are already well established in climate model evaluation. For example, satellite data have been used to assess climate models with regard to sea surface temperature (e.g. Jha et al., 2014; Lauer et al., 2017), sea ice extent (e.g. Turner et al., 2013; Shu et al., 2020), air temperature (e.g. Tian et al., 2013), cloud cover (e.g. Cesana & Waliser, 2016; Vignesh et al., 2020), and precipitation (e.g. Liu et al., 2012; Mehran et al., 2014).

With time advancing, also geodetic observational records have reached a length that allows for a comparison with climate models. Since 2002 the Gravity Recovery And Climate Experiment (GRACE, Tapley et al., 2004) and its Follow-On (GRACE-FO, Kornfeld et al., 2019; Landerer et al., 2020) satellite missions have been measuring the time-variable gravity field of the earth. Changes in gravity are caused by mass changes, and over the continents these changes are mainly caused by changes in terrestrial water storage (TWS), consisting of soil moisture, snow, groundwater, and surface water. As such, TWS is part of the global water cycle. It is one of the four variables needed to close the terrestrial water budget: temporal changes in terrestrial water storage are the net effect of precipitation counteracted by evapotranspiration and surface and sub-surface runoff. Climate change impacts on the water cycle manifest themselves as changes to the water budget equation, and TWS as an integral signal is sensitive to changes in any of its component storages (soil moisture, snow, groundwater, surface water) and fluxes (precipitation, evapotranspiration, runoff). Due to this holistic character, TWS is an important factor for the global water cycle. Acknowledging this crucial role, TWS was recently recognized as a new Essential Climate Variable (ECV) by the Global Climate Observing System (GCOS) Steering Committee.

Despite its importance in the climate system, which was already documented in many climate-related studies, TWS has hardly ever been used for the evaluation of coupled climate models (which are employed for long-term climate projections). This thesis investigates in which way TWS derived from satellite gravimetry observations can be used to assess the quality of coupled climate models regarding their representation of land water storage-related variables.

1.2. State-of-the-art and research gap

In this thesis, climate models are evaluated by means of space gravimetric observations from the GRACE and GRACE-FO missions. Over the last 19 years, these satellite missions have significantly contributed to understanding past climate change (Tapley et al., 2019). Due to their sensitivity to mass changes in the Earth system, their data were widely used in studies on the redistribution of water masses in the continental hydrosphere, in the ocean, and in ice covered regions. For example, mass variations for the Greenland and Antarctic ice sheets and mountain glaciers were derived from GRACE/GRACE-FO data (Sasgen et al., 2012; Velicogna & Wahr, 2013; Jacob et al., 2012). In the context of ocean studies, they help to interpret the global sea level record by determining the mass component of sea level change (Rietbroek et al., 2016; Chambers et al., 2017), and to assess ocean circulations (Landerer et al., 2015; Koelling et al., 2020). The derivation of terrestrial water storage changes from GRACE and GRACE-FO data lead to a variety of hydrological applications: inter-annual and long-term mass variations for numerous river basins were studied (e.g. Lettenmaier & Famiglietti, 2006), and anthropogenic ground water abstractions were assessed for the first time from space (Rodell et al., 2009; Tiwari et al., 2009). Also, the contribution of land water changes to sea level (Jensen et al., 2013; Reager et al., 2016) and the response of TWS to El Nino Southern Oscillation (ENSO) (Phillips et al., 2012; Ni et al., 2018) were investigated. Furthermore, GRACE contributes to drought and flood monitoring (Houborg et al., 2012; Reager et al., 2014), and was used to quantify return frequencies of extreme events in land water storage (Kusche et al., 2016). Recently, GRACE-derived TWS has even been shown to be strongly linked to variations in the atmospheric CO_2 growth rate (Humphrey et al., 2018). In addition to pure observation-based climate studies, GRACE/GRACE-FO data also contribute to model development: they were utilized to validate and calibrate both global hydrological models (Döll et al., 2014; Güntner, 2008; Lenczuk et al., 2020) and land surface models (Scanlon et al., 2018; Zhang et al., 2017). In recent years, increasing effort was also made in the combination of models and observations by assimilating GRACE data into hydrological models (Eicker et al., 2014; Khaki et al., 2017).

This thesis explores a new application of GRACE/GRACE-FO data in the evaluation of climate models. In the most general sense, a climate model is a computer program to calculate the earth's climate for the past, present, and future. Climate models exist in different complexities and include different components of the Earth system. A specific class of climate models are General Circulation Models (GCMs) that represent dynamical processes in the atmosphere, ocean, cryosphere, and land surface in a three-dimensional grid over the globe, and interactions between these domains. GCMs that additionally incorporate biogeochemical cycles (especially the carbon cycle) and allow for feedback of these processes to the physical circulations are called Earth System Models (ESMs). In this thesis we investigate global interactively coupled ESMs. This means that the input data for the different model components are generated internally within the model and are exchanged by means of the implemented feedback mechanisms. In particular, the coupled ESMs analyzed in this thesis do not assimilate any observations but evolve freely over time after initialization, restricted only by very few external data (e.g. the Sun's energy and CO_2 concentrations). The consequence of these free evolutions of model runs is that they reproduce climate variability in a statistical sense only, but not the exact timing of particular events. This represents a challenge for the comparison of model results to observations, and is a major topic of this thesis. The terms ESM and climate model are used synonymously throughout the thesis, referring to the above definition.

ESMs have frequently been evaluated for several climate variables using different types of observations. Especially, model skills in terms of sea and air temperature, and precipitation were studied for two different types of ESM forecasts: century-long climate projections (e.g. Su et al., 2013; Sillmann et al., 2013; Kim et al., 2020) and short-term decadal predictions (e.g. Corti et al., 2012; Mehrotra et al., 2014; Boer et al., 2019a). However, only very few studies to date have assessed ESM skills with respect to water storage-related variables. A regional study in the Mississippi Basin compared the annual cycle of GRACE TWS with output from nine ESMs to verify improvements in a newer model generation (Freedman et al., 2014). A large ensemble of different runs from one individual ESM was investigated by Fasullo et al. (2016) in order to attribute trends in GRACE to interannual variability or anthropogenic climate change. Climate-driven changes in precipitation as simulated by ESMs were also utilized by Rodell et al. (2018) to identify different drivers of GRACE TWS trends. Decadal climate predictions from ESMs have only been evaluated in an initial study comparing GRACE TWS to decadal model outputs from a single ESM using a rather simple metric (Zhang et al., 2016).

The aforementioned studies, which used GRACE data to assess ESMs, were either restricted to a specific region or to one individual climate model. So far, an extensive and detailed global-scale evaluation of multiple ESMs regarding their representation of land water storage is lacking, although the GRACE/GRACE-FO record of nearly 20 years provides a unique global data set to constrain total TWS and its variability. This thesis makes a first effort to comprehensively and thoroughly investigate if and how space gravimetric observations can be used to evaluate global coupled Earth System Models. The previous lack of such studies may also be due to the fact that a ″simple″ one-to-one comparison of observed and modeled TWS time series is not feasible, but the characteristics of both, observations and models, as well as their uncertainties have to be carefully considered before obtaining meaningful results. This thesis explores and documents ways to use GRACE and GRACE-FO data for climate model skill assessment, and is meant to enhance mutual benefits of climate modeling and geodesy by bringing models and observations together.

1.3. Outline

This doctoral thesis has a cumulative form. It consists of three published scientific papers (Appendix A) and a framework text, which highlights the research demand for the study, provides the thematic context of the publications and a gives a summary of their main outcomes. In order to point out the challenges arising from the comparison of gravimetric observations with climate model output, background information on both, global observations of terrestrial water storage and on coupled Earth System Models, are given in Chapter 2 and 3. In view of the differences between observations and models, several research objectives arise, which are formulated in Chapter 4. These objectives are dealt with in detail in the three papers included in the Appendix A. In Chapter 5 the main results of the papers are discussed in compact answers to the proposed research questions. The thesis′ main findings are summarized in Chapter 6 and an outlook is given on future research options.

1.4. Publication overview

The three scientific papers of this thesis deal with different aspects of climate model evaluation by means of space gravimetric data, which are schematically illustrated in Figure 1.1. A further explanation of the different topics covered by the papers is provided below in a short summary:

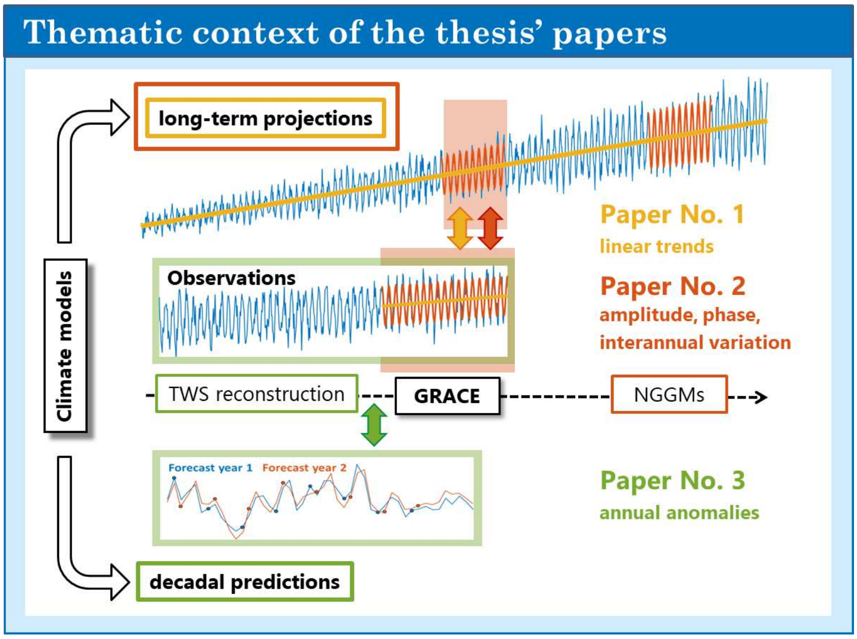


Figure 1.1.: Schematic illustration of the context of the thesis' papers.

Paper No. 1

Climate change may cause long-term wetting or drying in various regions of the world, reflected in terrestrial water storage trends. The identification of such regions is of substantial importance for water resources management. However, current ESMs used for climate projections until the end of the century are still quite discordant regarding long-term trends in land water storage-related variables. In Jensen et al. (2019, Appendix A.1, referred to as Paper No. 1) long-term, i.e. over more than two centuries, water storage trends from 21 climate models are evaluated against observed trends in TWS obtained from 14 years of data from the GRACE satellite mission. This is complicated, because (1) interannual climate variability masks long-term trends in the rather short observational time series, and (2) TWS from models and GRACE do not represent the same physical entity everywhere. To consider (1), we perform numerical model investigations to quantify the degree to which 14-year trends can be expected to represent long-term trends. To account for (2), we focus only on regions where climate models are largely concordant about the direction of long-term trends, and identify areas where conceptual discrepancies between modeled and observed TWS may be particularly large. This allows for a classification into areas where (a) climate-related TWS changes are supported by the direction of GRACE trends, (b) the mismatch of trends hints at potential for model improvement, or (c) interannual variations and/or human impact prevent reliable conclusions.

Paper No. 2

In addition to long-term wetting and drying trends (the focus of Paper No. 1), climate change will also affect the seasonal cycle and interannual variations of the terrestrial water cycle over the next decades. In Jensen et al. (2020a, Appendix A.2, referred to as Paper No. 2) climate model output is used to assess possible future changes in TWS variability with respect to present-day observed conditions to help defining scientific requirements for future satellite gravity missions. For this, best estimates for the annual amplitude, phase, and interannual variability of TWS from an ensemble of 17 climate models are derived, which are compared to GRACE observations in the time span 2002 - 2020 (the GRACE/GRACE-FO period). We find an overall reasonable fit of GRACE and models and discuss regional over- or underestimation by the models with respect to the observations. Afterwards, from the multi-model ensemble we derive global maps of the magnitude of changes in amplitude, phase, and interannual variability until 2100 and analyze the agreement of the models regarding the sign of these changes. Comparing these maps to accuracy maps derived from GRACE uncertainties, we estimate to which extent changes in the seasonal cycle could be detected after 30 years with current and future gravity missions. Furthermore, to serve as input for performance studies of future gravity missions, we select a specific model run that closely matches both GRACE observations and the best estimate from the model ensemble.

Paper No. 3

In Papers No. 1 and 2 long-term projections of climate models until the end of the century are analyzed using measures of TWS variability (seasonal cycle, interannual variations, linear trend). However, a dedicated one-to-one comparison of model and observational time series can only be realized for another category of climate model simulations, which are used to predict climate conditions for one decade (decadal predictions). In Jensen et al. (2020b, Appendix A.3, referred to as Paper No. 3) we assess the predictive skill of land water storage in an ensemble of climate models based on satellite gravity measurements. As the overlap time span between decadal predictions and GRACE is too short for deriving robust results, we resort to a century-long global reconstruction of TWS that relies on GRACE data (GRACE-REC) as a proxy for observed TWS. By means of GRACE-REC we globally investigate for how many years into the future annual TWS anomalies from decadal predictions fit better to the observations than those from long-term projections. For a better interpretation of the results, we also perform a regional skill assessment for different climate zones, and even on grid cell level. Furthermore, by comparing two successive model generations, we evaluate if the predictive skill and/or the reliability of the predictions improve for the new model versions.

2. Global Observations of Terrestrial Water Storage

Gravity measurements from the GRACE and GRACE-FO satellite missions are the observational basis for the evaluation of climate models in this thesis. Section 2.1 provides details on the mission concept, published data products and the derivation of water mass changes on the continents from gravity measurements. Furthermore, specific post-processing steps (applied in Paper No. 1 and 2) and the treatment of superimposed signals that obscure terrestrial water storage in the GRACE observations are outlined. Climate model simulations can benefit from the development of Next Generation Gravity Missions succeeding the GRACE-FO mission (Section 2.2, and Paper No. 2). Reconstructions of terrestrial water storage are valuable to extend the GRACE record prior to the mission launch time, rendering the evaluation of specific climate model experiments possible (Section 2.3, and Paper No. 3).

2.1. GRACE and GRACE-FO

The Gravity Recovery And Climate Experiment (GRACE, Tapley et al., 2004), a satellite mission jointly operated by the National Aeronautics and Space Administration (NASA) and the German Aerospace Center (DLR), was launched on March 17, 2002 and provided science data for more than 15 years until June 2017. Since May 22, 2018 its Follow-On (GRACE-FO, Kornfeld et al., 2019; Landerer et al., 2020) mission, a joint operation by NASA and the German Research Centre for Geosciences (GFZ), has been continuing the observations. The main objective of the missions is the determination of the earth's gravity field and its temporal variations.

The mission concept is based on satellite-to-satellite tracking, which means that two identically constructed satellites orbit the earth at a slowly decaying altitude of initially 500 km with an inter-satellite distance of about 250 km (Figure 2.1) and continuously determine this distance very accurately (relative measurement accuracy in the order of 10^{-12}) with a K-band microwave ranging instrument. The gravitational attraction of the earth, which is spatially and temporally variable, causes changes in the distance between the satellites while they are orbiting the earth. For example, when the satellite pair is approaching a positive mass anomaly (i.e. a stronger gravitational attraction), the first satellite is accelerated a little earlier than on the second, thereby increasing the inter-satellite distance. These metric distance changes (range-rates) are related to the gravity anomalies. Repeated observations of gravity anomalies from month to month, or year to year, allow to infer time-variations in the underlying mass distributions that are related to water storage changes or other mass transport processes in the Earth system.

For technology demonstration, GRACE-FO additionally carries a laser-ranging interferometer (LRI), further improving the range-rate accuracy by at least a factor of ten (Abich

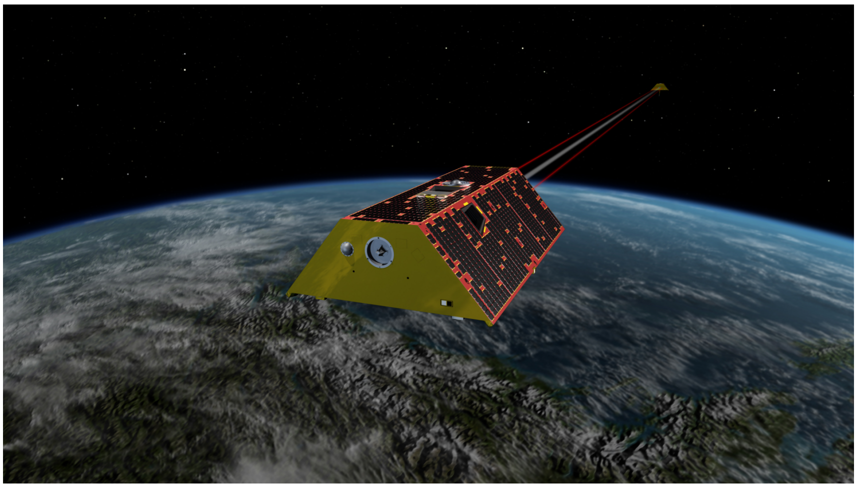


Figure 2.1.: Illustration of the GRACE-FO satellites in orbit. Source: NASA/JPL-Caltech, https://gracefo.jpl.nasa.gov/resources/41/grace-fo-in-orbit-view-3/ (visited 2020/14/12)

et al., 2019). Furthermore, each satellite is equipped with a three-axis accelerometer to capture non-gravitational forces, and two (three in case of GRACE-FO) star cameras to determine the attitude of the satellite, which is needed to relate the ranging data to the centre of mass of the satellites. Orbit determination is realized with Global Navigation Satellite System (GNSS) receivers. Due to frictional forces of the residual atmosphere and solar radiative pressure, the altitude of the satellite slowly decreases, causing an irregular ground track that leads to occasional periods of short repeat cycles.

With GRACE and GRACE-FO gravity field determination is possible with a spatial resolution of a few hundred kilometers from data accumulated over 30 days. Temporal changes of the gravity field are caused by mass variations in the Earth system, which can often be associated to climate processes. Mass variations mainly originate from the redistribution of water, as in ocean and atmosphere dynamics, continental hydrology, ice mass changes, and sea level variations, or due to mantle and crust dynamics, as glacial isostatic adjustment (GIA), and earthquakes. GRACE and GRACE-FO provide unique observations of mass transport divergence, i.e., the net inflow of mass to a certain location. With increasing length of the data record, they also start to valuably contribute to climate change studies (Tapley et al., 2019). This thesis demonstrates the feasibility of the evaluation of coupled Earth System Models (ESMs) as a new application for the GRACE and GRACE-FO data record.

2.1.1. GRACE and GRACE-FO gravity field solutions

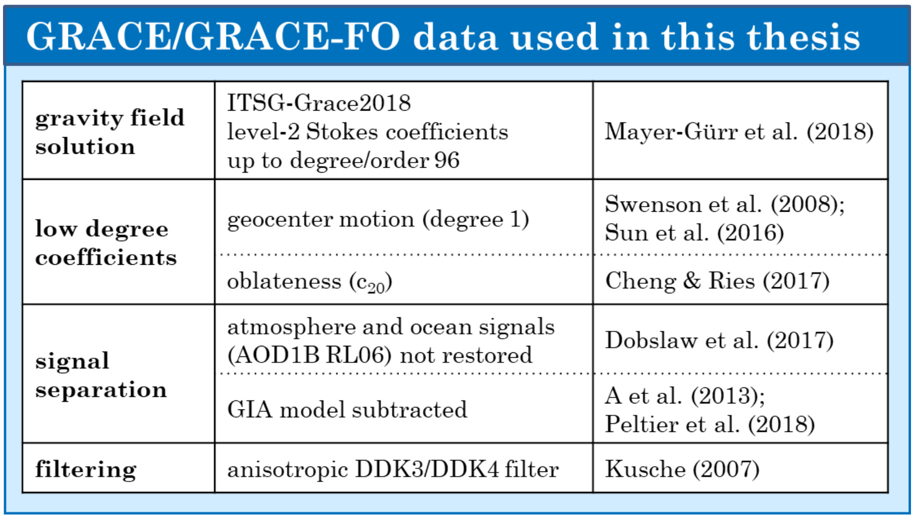
Raw data from the GRACE and GRACE-FO missions are processed by several analysis centers. In addition to the three official analysis centers, i.e., the GFZ, the Center for Space Research at the University of Texas (CSR), and the Jet Propulsion Laboratory at the California Institute of Technology (JPL), various other scientific institutes provide pre-processed GRACE data, e.g. the Institute of Geodesy at the Technical University of Graz, Working Group Theoretical Geodesy and Satellite Geodesy (ITSG). Furthermore, joint products combining data from different analysis centers are processed by the

Combination Service for Time-variable Gravity Fields (COST-G) of the International Gravity Field Service (IGFS).

The analysis centers apply different approaches to derive gravity field solutions from the raw GRACE/GRACE-FO observations. Usually, the parameters of the gravity field are estimated in form of monthly averaged spherical harmonic coefficients of the gravitational potential (Stokes coefficients) by means of a least squares adjustment. During data processing, tidal and non-tidal atmosphere and ocean mass variations and earth tides are subtracted from the observations by means of background models to avoid temporal aliasing effects caused by sub-monthly mass variations. Since these background models are not free of errors, GRACE gravity fields may contain residual tidal and non-tidal ocean and atmosphere mass changes. While the uncertainties of background models are difficult to quantify, they currently limit the accuracy of GRACE and GRACE-FO monthly solutions (Flechtner et al., 2016). Therefore, improvement of background models is one of the major challenges in the satellite gravimetry community.

The estimated Stokes coefficients (together with accuracy estimates) are published as monthly Level-2 data, that are freely and publicly available. They contain, besides the static gravity field, mass signals of glaciers and ice sheets, continental hydrology and the solid earth dynamics. The oceanic and atmospheric mass changes that were removed during processing can be restored by adding back the corresponding background data sets published together with the gravity field solutions. In this thesis, GRACE and GRACE-FO Level-2 monthly solutions from the ITSG-Grace2018 gravity field model (Mayer-Gürr et al., 2018) are employed. Details on the data and post-processing choices (Section 2.1.3) are given in Table 2.1.

Table 2.1.: Details on the GRACE/GRACE-FO data (and their post-processing) used in Paper No. 1 and 2. Cf. Section 2.1.1 for information on the gravity field solution and Section 2.1.3 for further explanation of the post-processing steps.



GRACE/GRACE-FO data used in this thesis

gravity field solution	ITSG-Grace2018 level-2 Stokes coefficients up to degree/order 96	Mayer-Gürr et al. (2018)
low degree coefficients	geocenter motion (degree 1)	Swenson et al. (2008); Sun et al. (2016)
	oblateness (c_{20})	Cheng & Ries (2017)
signal separation	atmosphere and ocean signals (AOD1B RL06) not restored	Dobslaw et al. (2017)
	GIA model subtracted	A et al. (2013); Peltier et al. (2018)
filtering	anisotropic DDK3/DDK4 filter	Kusche (2007)

2.1.2. From gravitational potential to surface mass changes

The Level-2 ITSG-Grace2018 gravity field solutions as used in this thesis are provided in terms of Stokes coefficients of the earth's gravitational potential. From these, global grids of mass changes have to be derived, which subsequently can be compared to climate model output. In the following the theoretical background for the computation of mass changes from gravitational potential is briefly outlined. More details on the potential theory is given in, e.g., Heiskanen & Moritz (1967).

Outside the earth, its gravitational potential V satisfies Laplace's equation $\Delta V = 0$. Therefore, it can be expanded into an infinite series of spherical harmonic functions

$$V(\lambda,\theta,r) = \frac{GM}{R}\sum_{n=0}^{\infty}\left(\frac{R}{r}\right)^{n+1}\sum_{m=-n}^{n} c_{nm}Y_{nm}(\lambda,\theta), \tag{2.1}$$

where λ, θ, and r denote the spherical polar coordinates, G is the gravitational constant, M and R are the mass and the reference radius of the earth, c_{nm} are the Stokes coefficients, and $Y_{nm}(\lambda,\theta)$ are the surface spherical harmonic functions defined as follows:

$$Y_{nm}(\lambda,\theta) = P_{n|m|}(\cos\theta)\begin{cases}\cos|m|\lambda & \text{for } m \geq 0\\ \sin|m|\lambda & \text{for } m < 0\end{cases} \tag{2.2}$$

The $P_{n|m|}$ are the fully normalized Legendre functions of degree n and order $|m|$.

The direct problem of potential theory, to compute the gravitational potential from a given mass distribution, can be solved uniquely. In contrast, the solution of the inverse problem of potential theory, to derive the mass distribution from a given gravitational potential, is non-unique. This is due to the fact that the same gravitational field can be generated by different mass distributions. To overcome this issue, the assumption is made that mass variations only occur in an infinitely thin layer at the surface of the earth. This assumption is approximately fulfilled when considering mass transport processes at centennial time-scales and shorter. Whereas mass redistribution in the interior of the earth happens on time scales of thousands of years, short-term mass variations are limited to the relatively thin layer between the floor of the ocean and the lower atmosphere layers. Instead of considering the total gravitational potential we focus on potential variations by subtracting the (time-)mean gravity field. The mass change induced by a surface potential change is represented by the surface density variation $\Delta\kappa(\lambda,\theta)$ which can also be expressed in spherical harmonics:

$$\Delta\kappa(\lambda,\theta) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}\Delta\kappa_{nm}Y_{nm}(\lambda,\theta) \tag{2.3}$$

The spherical harmonic coefficients $\Delta\kappa_{nm}$ of Eq. 2.3 are linked to the changes of the Stokes coefficients Δc_{nm} via

$$\Delta c_{nm} = \frac{(1+k'_n)}{(2n+1)}\frac{4\pi R^2}{M}\Delta\kappa_{nm}. \tag{2.4}$$

With the factor $(1+k'_n)$ two effects of mass change on the gravitational potential are considered: the first term accounts for the direct effect, the second term considers the indirect effect, which is the change in gravitational potential due to elastic deformation of the solid earth in response to mass changes at the surface. The effect of elastic deformation is quantified by the Load Love Numbers k'_n (Lambeck, 1988).

Combining Eq. 2.3 and 2.4, surface density variations $\Delta\kappa(\lambda, \theta)$ can be derived from given Stokes coefficient changes Δc_{nm} via

$$\Delta\kappa(\lambda, \theta) = \frac{M}{4\pi R^2} \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \frac{(2n+1)}{(1+k'_n)} \Delta c_{nm} Y_{nm}(\lambda, \theta). \tag{2.5}$$

For the sake of convenience, mass variations are usually expressed in equivalent water heights (EWH) rather than in surface density variations. EWH describe the height of water distributed uniformly over the surface that corresponds to the respective mass change. They are obtained by normalizing the surface density variation by the density of water $\rho_w = 1000\frac{kg}{m^3}$:

$$ewh(\lambda, \theta) = \frac{M}{4\pi R^2 \rho_w} \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \frac{(2n+1)}{(1+k'_n)} \Delta c_{nm} Y_{nm}(\lambda, \theta). \tag{2.6}$$

Equation 2.6 is the basic formula employed in this thesis to compute global grids of mass changes from GRACE/GRACE-FO Level-2 data.

2.1.3. Post-processing of GRACE and GRACE-FO data

To obtain mass changes from equation 2.6, the spherical harmonic coefficients provided by the analysis centers have to be post-processed. The commonly applied post-processing steps are described below. Details on practical choices for post-processing of the ITSG-Grace2018 gravity field solutions in Paper No. 1 and 2 are given in Table 2.1.

GRACE is insensitive to geocenter motion (the relative movement of the center of mass w.r.t. the center of figure), because the satellites circle around the center of mass. Therefore, the degree 1 Stokes coefficients, which are proportional to geocenter motion, cannot be determined from GRACE. The omission of degree 1 coefficients can have a significant impact on mass change estimates (Chen et al., 2005). Thus, they have to be replaced by an external time series, which can either be derived from satellite laser ranging (Cheng et al., 2010) or approximated by combining GRACE/GRACE-FO data and numerical ocean model output (Swenson et al., 2008; Sun et al., 2016).

Furthermore, the c_{20} Stokes coefficient, representing the earth's oblateness, exhibits a large uncertainty in the GRACE and GRACE-FO solutions. Therefore, it is usually replaced by a time series derived from satellite laser ranging (Cheng & Ries, 2017). Some processing centers also recommend to replace other low-degree spherical harmonic coefficients, for example all degree 2 coefficients (Dahle et al., 2019) and c_{30} (Cheng et al., 2011; Loomis et al., 2019).

After removing a static gravity field (e.g. the temporal mean of each c_{nm}) from the monthly solutions to obtain variations Δc_{nm}, often the effect of glacial isostatic adjustment (GIA) is reduced by a model. This only affects the linear trend in the time series of monthly solutions, as GIA acts as a linear increase or decrease of gravity due to the relaxation of the earth's crust and mantle after deglaciation since the last glacial maximum.

Due to the fact that the GRACE satellites observe the gravity field in an altitude of 400 - 500 km above the surface, especially the noise in the Stokes coefficients representing small spatial scales is amplified during downward continuation, leading to a small signal-to-noise ratio on these spatial scales. Furthermore, undersampling of unmodeled short-periodic changes of the gravity field (residual atmosphere and ocean mass changes) causes temporal aliasing effects. As the satellites move on a near-polar orbit and range-rates are measured along-track, the resulting spatial error pattern of GRACE monthly solutions is very anisotropic, characterized by north-south oriented stripes. Therefore, the monthly solutions have to be filtered in order to extract the signal. Advanced filter designs that account for the anisotropic error structure were developed (Swenson & Wahr, 2006; Kusche, 2007), and are still further improved by also considering temporal variations of the error structure (Horvath et al., 2018). Spatial averaging during filtering does not only effectively reduce the noise but also blurs the signal, causing so-called spatial leakage effects: neighboring signals are smeared into each other and therefore cannot be distinguished anymore in the filtered result. Depending on the application, leakage can have significant impact on mass change estimates (Baur et al., 2009, e.g.). Therefore, different strategies to mitigate leakage effects were developed and applied (e.g. Chen et al., 2015; Mu et al., 2017).

Eventually, the Δc_{nm} with degree 1 and other low degree coefficients replaced, filtered, and GIA reduced, are plugged into Eq. 2.6. Usually the processing centers stop the series expansion at a maximum degree of $n_{max} = 60$, or $n_{max} = 96$, or $n_{max} = 120$, corresponding to a spatial resolution s of about 330, 200, or 160km, linked by the relationship $s = \pi R / n_{max}$. Eq. 2.6 is evaluated for specific geographic locations λ, θ to obtain grids of mass changes in EWH. In the papers of this thesis, global 2° grids are computed from the Δc_{nm} of ITSG-Grace2018 with $n_{max} = 96$. Please note that due to filtering, neighboring grid cells in the 2° grids are highly correlated, so that the effective spatial resolution is less than the spatial sampling of about 222km at the equator.

Together with the Stokes coefficients the processing centers provide uncertainties for each monthly solution. For the ITSG-Grace2018 solutions full error variance-covariance matrices of the spherical harmonic coefficients are available. A unique feature of the ITSG gravity field estimation procedure is the introduction of an empirically estimated stochastic model into the least-squares adjustment, which applies a proper weighting scheme to different groups of observations and takes the temporal error correlation structure of the measurement noise and the background model uncertainties into account. As a result of this thorough stochastic modeling, the ITSG-Grace2018 a posteriori error estimations are considered as a close approximation of the actual error structure of the gravity field coefficients (Kvas et al., 2019). Therefore, to allow for a realistic assessment of the GRACE/GRACE-FO uncertainties on grid cell level in this thesis, we utilize the ITSG-Grace2018 error covariance matrices by performing strict variance propagation of Eq. 2.6, leading to grids of standard deviations of the respective $ewh(\lambda, \theta)$.

As an alternative to the spherical harmonic approach, in recent years also global mass concentration (mascon) estimates of the time-variable gravity field have evolved (Watkins et al., 2015; Save et al., 2016). Mascon solutions provide surface mass changes in equal-area grid cells that are computed by applying a regularization in the least-squares gravity inversion considering the error structure as well as geophysical signal covariances. Due to such spatiotemporal constraints the resulting solutions are not purely GRACE-based but have the advantage that leakage is reduced and no post-processing filtering has to be applied. Also Level-3 data (gridded data products) gain increasing attention, as the necessary post-processing was already applied here and therefore, they can be more easily used by a broader user community. Level-3 data products are for example provided by JPL[1] and GFZ[2] and can be explored in data visualization tools like the NASA GRACE Data Analysis Tool (DAT)[3] or the GFZ Gravity Information Service (GravIS)[4]. These products have the advantage of being ready-to-use for non-geodetic users, but lack of the possibility to apply a flexible post-processing and therefore are not tailored to specific (regional) applications.

2.1.4. Continental mass change signals in GRACE and GRACE-FO

The EWH grids derived from GRACE and GRACE-FO monthly solutions represent the integral signal of mass change on the surface and below for every location. Over the continents, these changes primarily originate from signals of terrestrial water storage (TWS), ice masses, residual atmosphere, residual GIA, and earthquakes. TWS itself is composed of soil moisture, snow, surface water, and groundwater. All these signals are superimposed, and with GRACE alone, separation of the different components is not possible. However, for a correct interpretation of GRACE-derived mass changes, signal separation is of crucial importance. Therefore, this issue is considered a major challenge in the GRACE/hydrology community and is subject of various ongoing projects (for example in the G3P project [5] or in Deggim et al., 2021).

In this thesis, GRACE/GRACE-FO data are compared to output from ESMs, which only provide total soil moisture content and surface snow amount as water storage-related variables. However, rigorous isolation of soil moisture and snow from the integral GRACE signal is difficult due to lack of data quantifying other sources of mass change and in view of the physics that govern the infiltration process at the soil surface. The thesis is therefore restricted to the discussion of the combined signal only. However, we identify regions where disturbing signals may have a particularly large influence on the results and mark these regions or mask them out prior to interpretation. Generally, we exclude Greenland and Antarctica from all our analyses, as mass changes in these regions are dominated by ice mass variations, which are not represented by the ESMs. In the following, the different signals possibly overlaying soil moisture and snow signals are briefly described and their treatment in this thesis is outlined:

Groundwater

Natural groundwater variability is implicitly contained to a large extent within the deeper soil layers of ESMs, because the water balance in the models is typically constrained and the mass transport to the ocean and atmosphere is limited (Liepert & Lo, 2013). Therefore, we abstain from separating natural groundwater variability from the integral GRACE signal when comparing to ESMs. However, anthropogenic groundwater depletion has to be considered, because this is not included in the ESMs. To estimate the magnitude of groundwater abstraction we make use of data from the hydrological model WaterGAP 2.2a (Döll et al., 2014). Net abstraction in WaterGAP is defined as groundwater with-

[1] grace.jpl.nasa.gov/data/monthly-mass-grids, last visit: 2/26/2021

[2] ftp://isdcftp.gfz-potsdam.de/grace/GravIS, last visit: 2/26/2021

[3] grace.jpl.nasa.gov/data/data-analysis-tool, last visit: 2/26/2021

[4] gravis.gfz-potsdam.de, last visit: 2/26/2021

[5] www.g3p.eu, last visit: 5/12/2021

drawals minus return flow from groundwater and surface water irrigation. Comparing the signal variability of the modeled groundwater time series to the total signal variability we identify regions largely influenced by anthropogenic groundwater depletion. We note that groundwater abstraction mainly occurs as a linear mass trend with weak year-to-year variations. Thus, it has only minor influence for studies focusing on annual, interannual or sub-seasonal signals, rather than linear trends.

Surface water

To estimate the effect of surface water storage we make use of an observational data set (Deggim et al., 2021) containing mass change time series of 283 large lakes and reservoirs derived from combining surface water levels (from satellite altimetry) with surface water extent (from remote sensing). In Deggim et al. (2021) this data set is actually applied to correct GRACE TWS time series for surface water storage. However, in this thesis we only use it to identify regions where the relative fraction of the surface water signal is large compared to the total signal.

Glaciers

To identify regions where glacier mass changes contribute to moisture dynamics, we access the Global Land Ice Measurements from Space (GLIMS) Glacier Database (GLIMS & NSIDC, 2005, updated 2020), currently containing the outlines of about 546300 glaciers. We use the polygons in this data set to identify grid cells that are covered by glaciers to a certain extent in order to mark them as regions probably affected by glacier mass changes. These regions make up about 3% of the land surface. We do not quantify actual glacier mass changes nor reduce them from the GRACE mass signals at this point.

Earthquakes

Earthquakes involving a substantial vertical mass displacement (normal fault or reverse fault earthquakes with a magnitude of about > 8.5) are detectable with GRACE (Han et al., 2013). In a first approximation the mass displacement causes a step function in the GRACE-derived time series of mass variations (with the discontinuity at the time of the earthquake), which overlays the TWS signal in earthquake regions. For a rigorous quantification of earthquake mass change signals (e.g. for the purpose of removal from the GRACE time series), co-seismic and postseismic effects have to be considered (Han et al., 2008). However, as we focus only on the identification of the maximum spatial extent in which TWS may be superimposed by earthquake mass variations, we neglect postseismic signals. Information on the spatial extent and magnitude of co-seismic mass variations contained in GRACE observations is, e.g., provided by Mayer-Gürr et al. (2018). They co-estimate the mass variations of the three largest earthquakes during the GRACE period (the Sumatra-Andaman 2004, the Chile 2010 and the Tohoku 2011 earthquake) together with the static gravity field model ITSG-Grace2018s.

Glacial isostatic adjustment

Even after the removal of a GIA model, residual GIA effects may remain in the linear trends derived from GRACE observations. The discrepancies between different GIA models are large, mainly due to incomplete knowledge about the ice history and mantle viscosity, and research on minimizing them is ongoing (Caron et al., 2018). There are also efforts to infer the mantle viscosity structure from GRACE data after disentangling hydrological and GIA mass trends (Steffen et al., 2010). However, to date it is difficult to distinguish (residual) GIA and hydrological signals in GRACE mass trend estimates. Therefore, care has to be taken when interpreting TWS trends in regions affected by GIA, as in Paper No. 1, where these areas are marked as uncertain regarding climate-related trends in GRACE. We note that the viscoelastic response of the mantle and crust to the past vanishing of ice covers can be approximated as a long-term linear effect, thus (residual) GIA does not affect annual, interannual or sub-seasonal signals.

The necessary post-processing of the data sets used for the consideration of the different signal components is described in the Supplementary Materials of the thesis' papers (Appendix A). The threshold determination to obtain masks with regions largely affected by mass signals other than soil moisture and snow is also delineated there.

2.2. Next Generation Gravity Missions

There are plans to continue the observation of the gravity field after the end of the GRACE-FO mission, which has a nominal lifetime of five years. The importance of a successor gravity field mission was highlighted by the most recent Decadal Survey for Earth Science and Applications from Space by the NASA (Committee on the Decadal Survey for Earth Science and Applications from Space et al., 2018), which listed mass change as a Designated Observable. The recent establishment of terrestrial water storage as a new ECV by the GCOS Steering Committee further amplifies the necessity of continuous mass change observations from space. Next Generation Gravity Missions (NGGMs) are currently being prepared to extend the GRACE and GRACE-FO record. Besides the mere extension of the time series, which would already have benefit for capturing climate change signals, also user requirements and needs for a higher spatial and temporal resolution and a higher accuracy for an NGGM have been identified. These user needs were collected among international experts of all relevant geoscientific application fields and summarized in several reports (IUGG, 2015; Pail et al., 2015; IGSWG, 2016).

From the identified user needs, threshold and target scenarios for the performance of a future gravity mission were derived. A mission fulfilling the threshold scenario would be clearly justified by significantly enhancing the current possibilities and enabling various new applications. Achieving the target scenario would mean a huge leap forward enabling the advance to completely new scientific questions (Pail et al., 2015). Generally, the threshold scenario requires an improvement of the performance by a factor of five compared to the GRACE mission. A further improvement by a factor of ten would be necessary for reaching the more ambitious target scenario requirements. Recently, these threshold and target requirements were also proposed in a joint NGGM Mission Requirements Document of the European Space Agency (ESA) and the NASA (European Space Research and Technology Centre, 2020). For hydrological applications the improvement of the spatial resolution is given higher priority than improvements in the temporal resolution or the latency of data availability. To realize a mission with an increased spatio-temporal resolution with higher accuracy, different NGGM concepts are currently being proposed. As the lack of spatial resolution of the GRACE mission mainly originates from its strictly north-south oriented orbit design with the necessity of extensive filtering, ground track patterns of alternative orbit designs were investigated (Murböck et al., 2014). Different promising concepts were identified: (1) a pendulum pair, where the two satellite orbits are slightly shifted against each other, (2) an in-line pair combined with a third satellite moving on a shifted orbit, (3) two in-line pairs with different inclinations, one on a polar orbit and one with an inclination of 65 – 70 degrees, the so-called Bender constellation (Bender et al., 2008).

To select a mission configuration and to demonstrate its potential value with respect to pre-defined user requirements, extensive end-to-end satellite simulations are typically performed (Pail et al., 2015). The main outcome of such numerical simulations is the rating of achievable performance of a mission concept regarding accuracy, spatial and temporal resolution, against the characteristics of the target signal. Several simulations have shown that enhanced satellite constellations, such as the Bender constellation, result in a significantly improved error structure and accuracy compared to the classical GRACE-type concept (Wiese et al., 2011; Daras & Pail, 2017; Purkhauser & Pail, 2019). So far, mainly short-term simulations covering the typical lifetime of a single satellite mission have been carried out. However, also long-term simulations over several decades are needed to assess the ability of missions to monitor climate variations such as long-term trends, interannual variability and changes in the annual cycle.

Simple error propagation from short-term simulations to long time periods does not provide adequate long-term performance estimates, because the relative error contribution of instrument errors and temporal aliasing errors (caused by short-period tidal and non-tidal signals) to the total error budget significantly changes with increasing averaging period. Full-fledged long-term NGGM performance simulations require realistic time series of the likely future evolution of TWS over several decades as input. Climate models can provide such realistic input time series from projections of climate conditions. Within this scope, the mutual benefit of satellite observations and climate models becomes apparent: on the one hand, continuous and extended time series of satellite observations are needed to validate and improve ESMs, on the other hand, realistic ESM output is needed to evaluate the performance of future satellite missions regarding long-term signals and trends.

2.3. Terrestrial water storage reconstruction

The GRACE/GRACE-FO time series has a length of about 19 years at the time of writing with a data gap of about one year between the two missions. While it is continuously growing and will hopefully be further extended by upcoming NGGMs, at this time the data record is still rather short for climate studies, which usually need time series of at least 30 years (as defined by the WMO). For some applications, e.g. the comparison of the annual cycle in climate models and observations, the limited length of the observational time series is rather uncritical, but especially for the direct comparison of observations with model results on time series level (as feasible for short-term decadal predictions), a longer data record is needed. Therefore, in Paper No. 3 we make use of estimates for TWS changes prior to the GRACE era that were reconstructed from meteorological quantities (Humphrey & Gudmundsson, 2019). By assuming that short-term anomalies of TWS are mainly driven by fluctuations in the relevant atmospheric drivers, Humphrey & Gudmundsson (2019) use precipitation and temperature data from atmospheric reanalyses to reconstruct past anomalies of TWS. The statistical model that links the meteorological data to TWS changes is derived from GRACE observations. The GRACE-REC data set covers the time span 1901 – 2014 and was shown to be close to the original GRACE observations within the overlapping period. GRACE-REC was also evaluated against independent data sets, indicating that it states a reliable proxy for water storage changes also for the years prior to the GRACE era. A drawback of the data set is that only TWS anomalies can be estimated, but no linear trends or annual cycles.

The GRACE-REC data set consists of an ensemble of 100 time series that slightly differ from each other, thereby representing the uncertainty of the reconstruction. The ensemble members were generated by employing a spatial autoregressive noise model, which reproduces the spatial and temporal autocorrelation structure of the empirical residuals. With the use of the ensemble it is possible to derive realistic aggregated errors for basin-averaged time series that also include correlations.

In Paper No. 3 the value of reconstructed TWS as a proxy for real TWS was shown by applying 40 years of GRACE-REC for the evaluation of short-term climate model predictions, which would not have been possible yet with the short GRACE record.

3. Coupled Earth System Models

The information obtained from climate models cannot be interpreted in the same way as observational time series. To make this clear, this chapter provides basic information on the structure of climate models, their results and their uncertainty (Sections 3.1, 3.2, and 3.3). From this, a distinction between climate projections and climate predictions arises, which demand different strategies for the comparison with observations (Sections 3.4, and 3.5). Furthermore, the representation of land water storage in climate models differs from the observed TWS, implying further challenges for the comparison (Section 3.6).

3.1. Structure of climate models

Coupled Earth System Models (ESMs), also referred to as climate models, are mathematical representations of the major components of the climate system (atmosphere, land, ocean) and their interactions in a computer programme (Gettelman & Rood, 2016). Each component and each process in a climate model relies on basic physical and chemical laws, which are Newton's classical physical mechanics, such as the conservation of momentum and mass, and the laws of thermodynamics, such as the conservation of energy.

The processes in the climate system are formulated in ordinary and partial differential equations, which are solved on a spatially discrete three-dimensional grid covering the whole earth. Each grid cell belongs to one of the components (atmosphere, land, or ocean) and contains different physical properties, e.g., temperature, water vapour content, pressure, which are obtained by solving the model equations (prognostic variables) and describe the state of the grid cell. The model grid cells can also exchange fluxes of energy, momentum and mass across their boundaries. The entity of all grid cell states characterizes the current state of the climate system. Other variables that are not directly contained in the model equations, e.g., precipitation, relative humidity, can be derived as diagnostic variables from the prognostic variables (Randall, 2017). However, as secondary variables, their determination might be less robust than for prognostic variables, leading to larger discrepancies between different models and highlighting the need for thorough validation.

To describe the spatial and temporal evolution of the grid cell states, the differential equations in the model are integrated in time starting from initial conditions by obeying certain boundary conditions as defined by the particular model experiment. For each time step, the current state for each grid cell is computed based on the last state considering all processes implemented in the model's equations. First, the prognostic variables are updated in each grid cell, then the fluxes between grid cells are computed from the updated states, altering the states again, and so on. Since each process in each grid cell

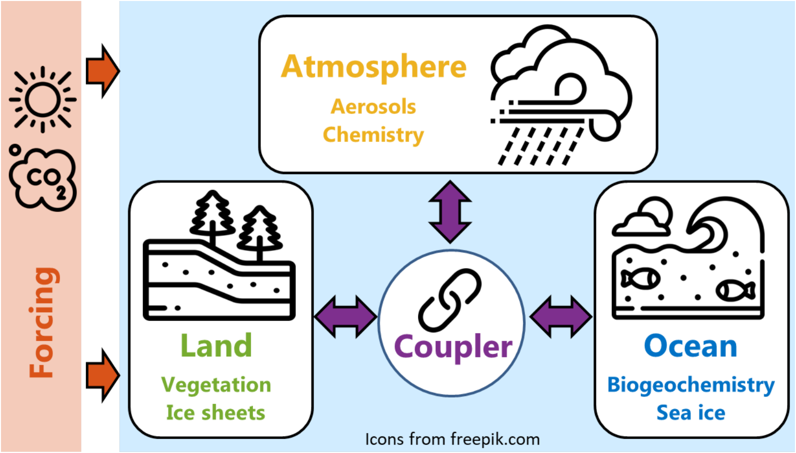


Figure 3.1.: Different model components interact by means of a coupler module. The main component models (atmosphere, ocean, land) contain sub-component models as for aerosols, chemistry or sea ice.

is always constrained by the basic principles of mass and energy conservation, the final outcome of the model is a realistic state of the climate system.

Most climate models are discretized (in very different ways) horizontally and vertically, to be able to represent horizontal and vertical motions and energy fluxes that mostly occur in a thin spherical shell at the surface of our planet. Spatial resolution is generally limited for globally discretized models, so that not all small-scale processes can explicitly be modelled. To mimic such effects, which are sometimes highly important for the evolution of the climate, some processes are parameterized. This means that a process is not exactly described by formulas, but the effect of the process on the state of the grid cell is approximated. The parameters for such processes are derived empirically from observations. The drawbacks of parametrizations are their dependence on the quality of the observations and that they are often ambiguous, meaning that different combinations of parameters can lead to the same outcome. Depending on the choice which processes are considered, which are modeled dynamically and which empirically, and how the parameters are chosen, the results differ from model to model, and hence climate change predictions differ.

Usually, a climate model consists of an atmosphere model, a land surface model, and an ocean model that can either be run independently or coupled together with a coupler module, which handles the exchanges between the different components along their spatial boundaries. A schematic overview of the components of a climate model and their interactions is given in Figure 3.1. The state of the climate does not only depend on its previous states, but is also largely influenced by external boundary conditions. For example, solar radiation is not constant over time, but varies over multiple cycles ranging from decadal to potentially multi-millennial time scales, and changes in solar radiation have a significant impact on the earth's climate. Furthermore, time-variable or even slowly increasing greenhouse gas (GHG) emissions (CO2, methane, nitrous oxides, and halocarbons) and aerosols (e.g. emitted during volcanic eruptions) alter the capability of the earth to absorb heat and therefore influence climate. Also land use changes contribute to climate change by altering surface albedo. Thus, to obtain realistic results from climate models, these forcings have to be provided as boundary conditions when running a model. The forcing data is either obtained from observations (for simulations of the past) or prescribed for the future in a certain forcing scenario. Opposed to land surface models or other single component models, in coupled ESMs the forcing data described above are the only observations constraining the model. All other drivers for individual sub-systems (e.g. precipitation determining soil moisture, atmospheric winds driving ocean currents, and other interactions) are calculated and exchanged via the coupler within the model itself.

To start a model run, initial conditions are required. These are usually taken from another model or from observations, or a combination of both. However, if these initial conditions are not thermal-dynamically balanced as the model equations require, the model will adjust its states in sometimes unrealistic oscillations until its preferred climatological balance is reached. During this so-called spin-up period, the model results are not reliable and have to be disregarded. Depending on the processes spin-up takes between some hours (for the atmosphere) and up to hundreds of years (for the deep ocean). The model states after a sufficiently long spin-up time are then used as the initial conditions for actual model experiments for scientific purposes. The role of initial conditions is particularly critical for short simulation runs over only a decade, which is discussed in Section 3.5.

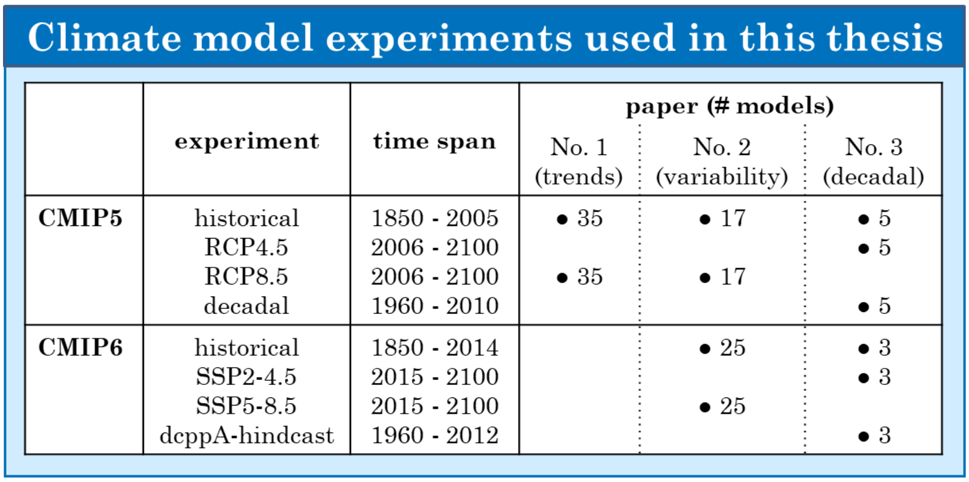
3.2. Coupled Model Intercomparison Project

In this thesis we exclusively use output from climate models taking part in the Coupled Model Intercomparison Project (CMIP). CMIP is an initiative of the World Climate Research Programme (WCRP) with the goal to improve knowledge on past, present and future climate changes by fostering the comparison and analysis of global climate models. This is achieved by the definition of a set of common experiments and the provision of forcing data sets to ensure comparability and enable assessment of model performance. Furthermore, CMIP features standardized model output and publicly available data access. The model results provided to CMIP serve as a basis for the Assessment Reports of the IPCC and therefore represent an important source of information for policy makers and stakeholders developing climate change adaptation and mitigation strategies.

The most recent IPCC Assessment Report (AR5, IPCC, 2013) is based on results from the 5th phase of CMIP (Taylor et al., 2011). Model results contributing to CMIP5 have been frequently evaluated in the last years, and are also the basis of Papers No. 1 and 3. Currently, the 6th phase of CMIP is running (Eyring et al., 2016), with 49 international modeling groups participating by successively making their model results available. CMIP6 results will be the basis of the next upcoming IPCC Assessment Report (AR6, currently due for release in 2022), and were employed in Paper No. 2 and 3. In the following, further information is specifically given for CMIP6.

Important experiments within CMIP6 that were used in this thesis (Table 3.1), are the historical (1850 – 2014) and the Shared Socioeconomic Pathway (SSP, 2015 – 2100 or 2300) scenario simulations (Riahi et al., 2017). Similar experiments are defined in CMIP5 with the historical (1850 – 2005) and the Representative Concentration Pathway (RCP, 2006 – 2100 or 2300) scenarios. The forcing data of the historical experiment are based on observations and include solar variability, (anthropogenic) GHG

Table 3.1.: CMIP5 and CMIP6 experiments used in this thesis. The numbers denote the number of different models that were utilized for the respective study.



Climate model experiments used in this thesis

	experiment	time span	paper (# models)		
			No. 1 (trends)	No. 2 (variability)	No. 3 (decadal)
CMIP5	historical	1850 - 2005	• 35	• 17	• 5
	RCP4.5	2006 - 2100			• 5
	RCP8.5	2006 - 2100	• 35	• 17	
	decadal	1960 - 2010			• 5
CMIP6	historical	1850 - 2014		• 25	• 3
	SSP2-4.5	2015 - 2100			• 3
	SSP5-8.5	2015 - 2100		• 25	
	dcppA-hindcast	1960 - 2012			• 3

emissions, volcanic aerosols and land use data sets. In the nine SSP scenarios different assumptions are made on the future demographic and economic development regarding sustainable and fossil-fueled energy consumption. They range from SSP1-1.9 with the assumption of extensive international collaboration, substantial reduction of GHG emissions, and limitation of the global temperature increase to below 1.5°C, to SSP5-8.5 assuming a "business as usual" with high fossil fuel consumption, low rate of renewable energies and minor collaboration between countries. To illustrate the impact of different SSP forcing scenarios on climate evolution, Figure 3.2 (taken from Cook et al., 2020) exemplarily shows the global average surface air temperature anomalies simulated with several CMIP6 models under four different SSP scenarios. To cover the GRACE/GRACE-FO time span, the CMIP historical and SSP simulation runs were concatenated. As shown in Table 3.1, in Paper No. 1 and 2, where long-term climate projections (over more than a century) were analyzed, we used the RCP8.5 and SSP5-8.5 runs. To date, this scenario seems a plausible assumption, as it shows more agreement to historical CO_2 emissions than other scenarios (Schwalm et al., 2020). For consistency with other studies, in Paper No. 3 the RCP4.5 and SSP2-4.5 runs were utilized. Please note that within the very first years of the different RCP/SSP scenarios, to which this study is limited, they do not significantly differ from each other.

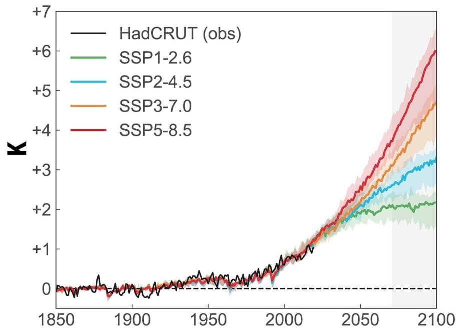


Figure 3.2.: From Cook et al. (2020): Global annual average surface air temperature anomalies for four SSP scenarios with ensemble mean (coloured lines) and observations (black).

Within the context of CMIP6 several other Model Intercomparison Projects (MIPs) are endorsed, for example the OMIP (Ocean), LS3MIP (Land Surface, Snow and Soil Moisture), and the DCPP (Decadal Climate Prediction Project), which address specific questions and augment the core experiments defined by CMIP6. In addition to long-term projections as the SSP experiments, especially results from the DCPP are considered within the scope of this thesis. Within the DCPP, modeling groups perform short-term simulations, so-called decadal predictions, running merely over 10 years from their initialization date. Such a near-term component was also included in the CMIP5 experimental design. In contrast to climate projections (e.g. SSP experiments), which depend on assumptions on the future behavior of humanity, the decadal predictions provide unconditional forecasts similar to weather predictions. The difference between projections and predictions is further elaborated in Sections 3.4 and 3.5.

3.3. Uncertainty in Earth System Models

The uncertainty of climate model projections arises from three different sources: initial condition uncertainty, scenario uncertainty and model uncertainty (Hawkins & Sutton, 2009), as discussed below.

- **Initial condition uncertainty (internal variability):** The initial state of the system influences the prediction of the following states to a certain extent. An example is the weather forecast: The weather tomorrow depends on the weather today. In the climate system, some processes are influenced by the initial conditions up to a decade ahead, mainly due to heat transfer in and from the deep ocean (Bryan et al., 2006). El Niño-Southern Oscillation (ENSO) phenomena are predictable from the current state for several months in advance. Also terrestrial water storage is assumed to have a memory of several years in many regions (Yuan & Zhu, 2018). Thus, initial condition uncertainties dominate the total uncertainty on time scales up to a few years but afterwards they only play a minor role in the error budget. The initial condition uncertainty can be quantified by assessing the internal variability when running a model several times with the same forcing but with slightly different initial conditions.
- **Scenario uncertainty:** Climate simulations are forced by external data, for example with the GHG emissions prescribed in the CMIP6 SSP scenarios. These forcing data sets are highly uncertain, as they depend on assumptions on the future human behavior, the societal and economic development. Whereas for some years in advance the demographic development and fossil/renewable energy consumptions are quite predictable, nobody can foresee how our societies will behave in 30, 50 or 100 years. Therefore, the scenario uncertainty grows with increasing lead time and is the dominating uncertainty for long-term climate projections.
- **Model uncertainty (structural and parameter uncertainty):** Model uncertainties arise from the imperfect representation of the complex processes in the Earth system. They depend on the choice of the model developers which processes to include in which way in the model, how to implement the interaction between processes and which processes to neglect. Model uncertainties can be split into structural and parameter uncertainty (Tebaldi & Knutti, 2007). Parameter uncertainties originate from the parametrization of processes that are too small to be explicitly represented in the model. They might be partly assessed by running the same model with varying parameters, but are in general difficult to disentangle. Structural uncertainties

include the choice of the processes to include and also the choice of the grid, the resolution and the numerical methods to solve the equations. They can only be quantified by jointly analyzing different models.

One strategy to quantify the different types of uncertainty is to run models multiple times, using slightly varying initial conditions, and thereby create an ensemble of model results, each representing a possible evolution of the future climate. Usually, different initial conditions are taken from different points in time of the spin-up run. From the variability of the runs of an individual model ensemble, the initial condition uncertainty can be obtained. The scenario uncertainty can be quantified by using multiple scenarios and comparing the ensemble results of the different scenarios, and the model uncertainty by running different models with the same scenario.

The latter results in a so-called multi-model ensemble, that is very valuable for assessments of possible future climate conditions and their likelihood. Multi-model ensemble analysis is broadly carried out for the results presented in the IPCC reports (Flato et al., 2013), where the multi-model mean is considered as a best estimate, which is more robust and less uncertain than a single model result. As a measure of uncertainty of such multi-model means, often the ensemble spread (the standard deviation of all model results around the mean) is provided. For many variables of the climate system the uncertainty of predictions and projections is still very large (Collins et al., 2013). Reducing the models' uncertainties is one of the major challenges faced by climate modelers.

While it is crucial to quantify the uncertainty of multi-model averages, it is not trivial and discussed in some detail by Knutti et al. (2010). If each model would be a completely independent representation of the real climate system, the uncertainty of the multi-model mean would decrease with the square root of the number of models averaged, and could be narrowed infinitely by adding more models. However, models are not independent, but partly share the same component models and parametrizations. The degree of dependence is difficult to quantify, and several strategies are pursued, which rely on assumptions ranging from full dependence to no dependence at all (Pennell & Reichler, 2011; Sanderson et al., 2015). In addition to the non-Gaussian distribution of model results, systematical biases hamper the assessment of realistic uncertainties. Due to our incomplete knowledge about processes in the climate system, omission of processes and interactions, and simplifications due to the model resolution, the model results are not necessarily distributed around the true value but can be systematically biased. Therefore, multi-model uncertainties might be overoptimistic and misleading for stakeholders regarding the reliability of models.

There is an ongoing debate about if and how models should be weighted in a multi-model mean, e.g. by downweighting models that fit less to observations than others (Knutti et al., 2017). However, weighting must be tailored to the specific applications, as the match of different models to observations largely depends on the investigated variable and on the region of interest.

3.4. Long-term projections

Coupled climate models are externally driven only by solar variability, GHG concentrations, aerosols, and possibly land use change. Other drivers (temperature, precipitation, wind, etc.) are not fed into the models from external observations but are generated by the different model components and exchanged between them as the simulation run progresses. This means that ESM time series are not constrained by real world conditions (besides the forcing described above) but develop freely from the date of their initialization. As the impact of initialization only lasts as long as the typical memory time-scale of a certain system, and long-term climate simulations are usually initialized in 1850, modeled time series of climate variables are not (meant to be) directly comparable to real observational time series in terms of the exact timing of troughs and peaks in their course. Due to the random nature of interannual and sub-seasonal variations climate models can only attempt to reproduce the statistical properties (i.e. magnitudes and frequencies) of the observations at the right place, but not a particular event at the right time.

As an example, in Figure 3.3 the modeled TWS time series of five different runs from a specific climate model (MPI-ESM1-2-LR) over 30 years are shown (top panel). The model runs differ only by their initial conditions at the starting point of the experiment in 1850 but are all created with the same forcing data set (historical and SSP5-8.5). The lower panels in Figure 3.3 display the different signal components, i.e. seasonal cycle, linear trend, interannual and sub-seasonal signal of each time series. Whereas the phase and amplitude of the seasonal cycle is rather homogeneous among the five model runs, the other components largely differ from run to run. This illustrates the effect of internal model variability and highlights that a direct comparison of model output and observations on time series level is not feasible.

For many applications (especially for long-term projections of climatic conditions as e.g. evaluated in the IPCC Assessment Reports), multi-model averages are built. Whereas this is useful for obtaining a "best estimate" for the long-term evolution of a climate variable and to assess the model spread, the year-to-year and sub-seasonal variabilities are largely smoothed out during averaging. The magnitude of the interannual and sub-seasonal variability in the multi-model mean time series is therefore much smaller than for individual model runs (black curve in lower two panels of Figure 3.3). Thus, also a multi-model average time series is not suitable to be directly compared to an observational record. In Papers No. 1 and 2, where long-term climate projections are analyzed, the challenges arising from their random nature are addressed.

The outcome of long-term climate projections depends on the prescribed scenario (e.g. the SSP scenario) that was chosen for the experiment. Therefore, conclusions on the future evolution of climate are conditional forecasts, in the meaning of "*if* GHGs and other forcings develop in this way, *then* climate will react in that way". This is very useful for assessments of the range of possible future conditions and therefore has a high societal relevance in terms of developing sustainable mitigation and adaptation strategies to climate change. However, long-term projections cannot provide concrete information on the year-to-year development of climate variables. For example, one cannot derive from climate projections whether or not a region affected by a drought will recover from it over the next years or if it will suffer from a more severe drought, simply because this information is random in climate projections.

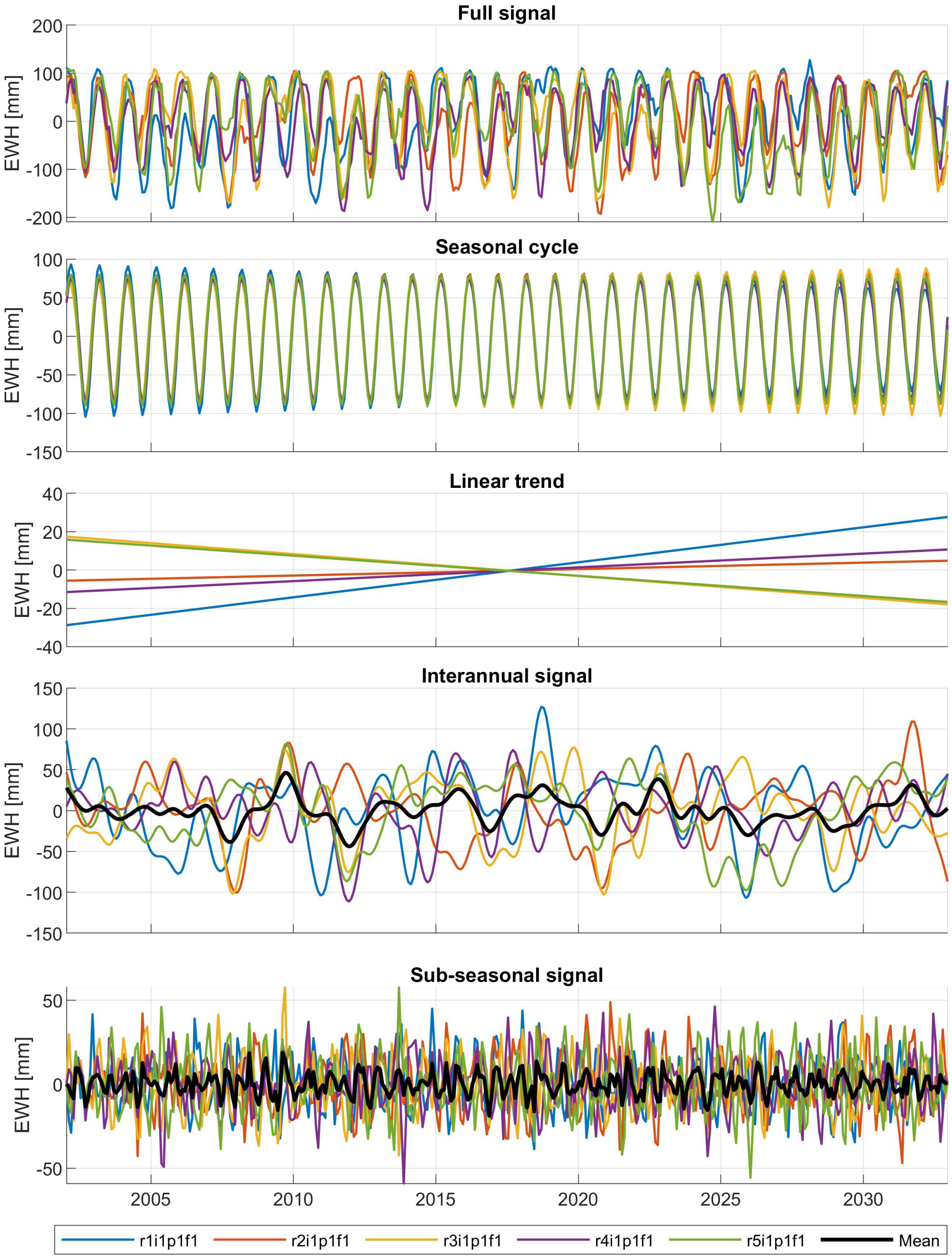


Figure 3.3.: Illustration of internal model variability: decomposition of modeled TWS time series into linear trend, seasonal, interannual, and subseasonal signal for different runs of a climate model (MPI-ESM1-2-LR). The location is 10°E and 53.5°N (Hamburg, Germany).

3.5. Decadal predictions

Reliable forecasts on year-to-year developments (such as drought recovery or intensification) would be extremely helpful for rather short-term decisions on water management and agricultural watering plans, but they cannot be provided by conditional forecasts such as long-term climate projections. In the last years, ESMs have been used for making unconditional forecasts for up to 10 years in advance by frequently initializing them with observations or reanalysis data.

Such unconditional forecasts are possible, because there is a certain time span in which the influence of initial conditions (internal variability) prevails over the model response induced from external forcings. Within this time span the climate system is deterministic in a sense that its future status depends on the past and present conditions, thereby making it potentially predictable. Model runs limited to this time span (up to a decade), are therefore called predictions, whereas model runs extending beyond this time, are called projections because they mostly depend on the prescribed forcing scenario. In Paper No. 3 decadal predictions of land water storage in CMIP models are analyzed. The following sections describe aspects of decadal predictions that have been considered but not been described in the paper for reasons of limited space.

3.5.1. Predictability in climate models

Decadal prediction is a relatively new field in climate modeling (Meehl et al., 2009, 2013), and it is still unclear to which extent climate is predictable and how well climate models are able to reproduce predictability. In principle, predictability is limited due to the fact that climate is very sensitive to tiny perturbations in the initial conditions leading to very different outcomes with growing forecast time. This effect is known in chaos theory as the butterfly effect (Lorenz, 1963). As it is impossible to exactly know all details of the initial status, virtually no system is perfectly predictable. The sensitivity to initial states, or "the rate of separation of initially close states" (Boer et al., 2019b), therefore characterizes predictability.

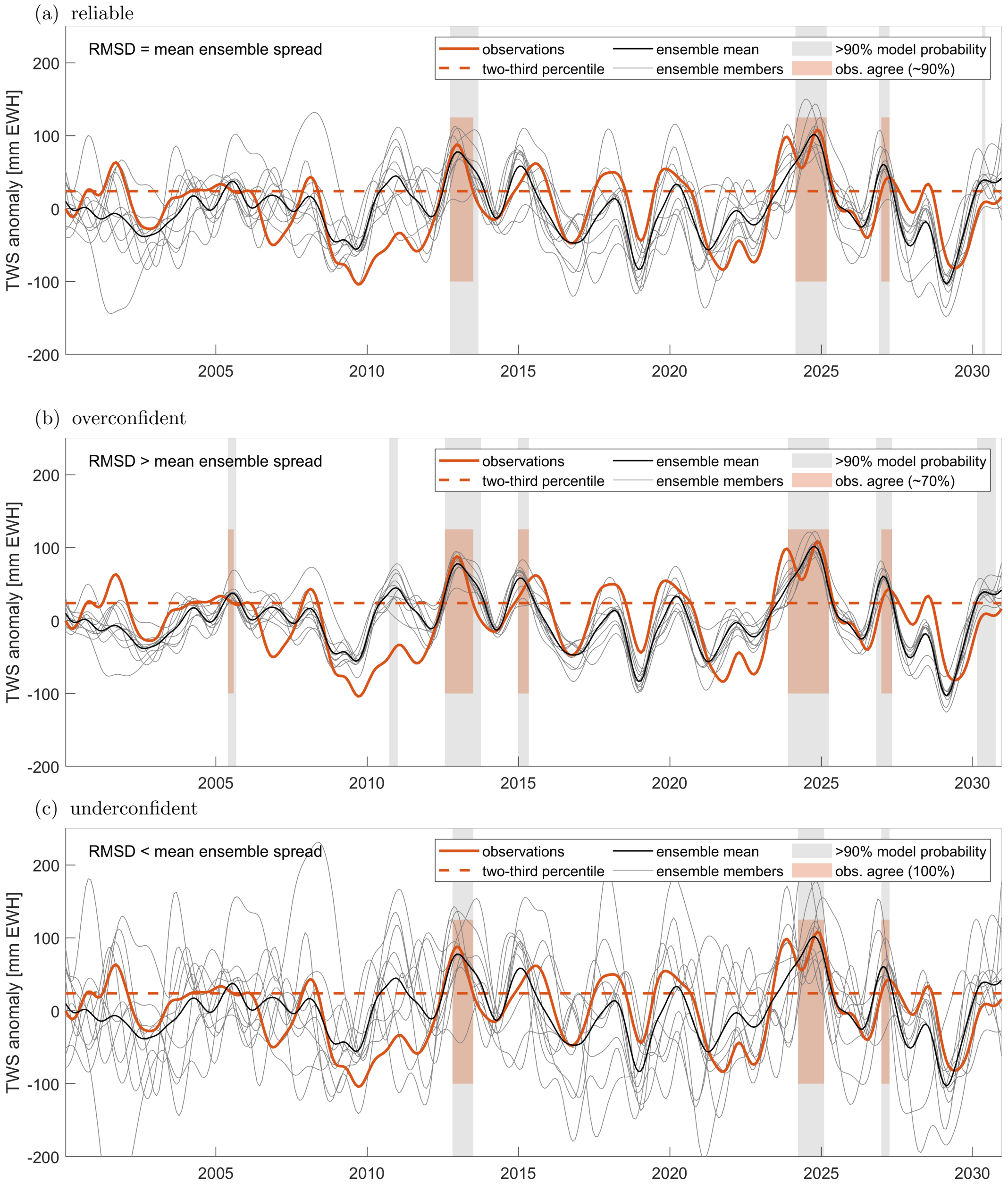
In decadal prediction, a distinction is made between *potential* predictability ("the ability of the system to be predicted") which depends on the inherent chaotic nature of the system, and *actual* predictability ("the ability to predict the system") which depends on the quality of climate models and their evaluation. The latter is assessed as the forecast skill by comparing model results to observations, e.g. using anomaly correlations, root mean squared deviation (RMSD) or other skill scores. Due to the fact that climate models are imperfect representations of the real climate system and observations are incomplete and uncertain, the actual skill does not correspond to the potential predictability. However, the potential predictability of the real system is difficult to characterize, because there is only one realization of it. Therefore, climate models are not only used to actually forecast future conditions but also to assess the potential predictability. This is done, e.g., by evaluating signal-to-noise ratios in model ensembles (Boer et al., 2013). Assuming that climate models are able to reasonably reproduce many aspects of the real system, also model-derived potential predictability should reasonably approximate real potential predictability. Usually, potential predictability is larger than actual predictability, and the difference indicates the potential for further improving the model (Boer et al., 2019b).

One goal in decadal prediction is to quantify the time span of predictability for different climate variables and thereby their potential value to be used for initialization of decadal predictions. Multi-year to multi-decadal variations in the Pacific and Atlantic ocean were identified as a major source of decadal climate variability (Keenlyside et al., 2008; Gulev et al., 2013; Wanders & Wada, 2015). Initializing the ocean component with observations therefore increases the predictive skill of ESMs for a variety of climate variables. So far, decadal prediction skill has mainly been demonstrated for sea surface and air temperature (Corti et al., 2012), and (moderately) for precipitation (Doblas-Reyes et al., 2013; Mehrotra et al., 2014), but also first studies showing skill for TWS were conducted (Zhang et al., 2016, Paper No. 3). Whereas the impact of ocean initialization has been established with these studies, a possible benefit of initialization of other components (sea ice, land surface, stratosphere) still has to be assessed (Bellucci et al., 2015). First model studies provide evidence that, due to its long memory of several years, initialization of TWS has potential to improve the skill of climate predictions (Yuan & Zhu, 2018). Furthermore, improved land surface components with deeper soil layers in ESMs would contribute to an improvement of predictive skill (Bunzel et al., 2018). There are indications that predictability derived from current decadal prediction ensembles is underestimated compared to real predictability (Eade et al., 2014; Smith et al., 2019).

3.5.2. Quality assessment of decadal predictions

The quality of decadal climate predictions (which are called hindcasts if run for the past) is assessed by comparing them to observations. Ideally, a forecast (or hindcast) should be accurate (fit to observations), skillful (provide valuable information) and reliable (events should actually occur with the forecast probability). These three properties, accuracy, skill, and reliability, were also assessed in Paper No. 3, but not explained in detail:

- **Accuracy:** a forecast is accurate if it matches the observations (e.g. in terms of correlation or RMSD).
- **Skill:** a forecast has skill if it is better (in terms of accuracy) than a reference forecast. This reference can be climatological values, i.e. the long-term averages derived from observations. It can also be persistent forecasts, which means to keep the forecast values constant to the initial values over the whole forecast period. Only if a forecast is better than such trivial forecasts, it is valuable for users. When evaluating decadal predictions it is common to compute the skill with respect to uninitialized projections, because this quantifies the added value from the initialization and provides information on the predictability of the investigated quantity.
- **Reliability:** A forecast system is reliable if the forecast probabilities match the probabilities of the observed event. Reliability can only be derived from an ensemble of hindcasts, because no probabilities can be derived from a single model run. The definition of reliability might become more intuitive in the following example (illustrated in Figure 3.4): We assume the event that a TWS value in a monthly time series is larger than the two-third percentile TWS value (dashed orange line). This two-third percentile threshold is derived from observations (orange curve) over a long time span (where ”long” can be defined in different ways, e.g. the hindcast time span). The climatological (that means averaged over a long time) probability of observing such an event is 33%. For each month of the hindcast



(a) reliable

(b) overconfident

(c) underconfident

Figure 3.4.: Schematic illustration of (a) reliable, (b) overconfident and (c) underconfident ensemble forecast systems for the event that a TWS value is above the two-third percentile (dashed orange line) of the observations (orange curve). Gray shaded areas mark the occasions where 90 - 100% of the ensemble members (gray curves) are above the two-third percentile. Parts of the gray shaded areas are marked in orange, when the observations are above two-third percentile as well. The RMSD is computed between observations and ensemble mean (black curve), the mean spread as the square root of the mean variance of the ensemble members around the ensemble mean.

period the percentage of ensemble members (gray curves) exhibiting a TWS value above this threshold is computed. This describes the modeled probability for this event for each month. Then, the modeled probabilities are compared to the observed probabilities: all months where the model probability lies in a certain probability bin (e.g. 90 - 100%, gray shaded areas in Figure 3.4) are compared to the actual observed TWS values at these particular months. If the observed values are actually above the threshold in 90-100% of these months (orange shaded areas), the forecast system is said to be reliable (at least for this probability bin). This relationship (forecast probability equals observed probability) should not only apply for the 90-100% probability bin but also for all other probability bins. E.g., in 40 - 50% of all months, where the model probability for the event (TWS above two-third percentile threshold) is 40 - 50%, also the observations should lie above this threshold.

This definition of reliability implies that a forecast can be perfectly reliable but still have no skill: in case of the event "TWS above two-third percentile" its climatological probability is 33% (i.e. by definition in 33% of the months the observations lie above this percentile). If the forecast system always predicts the climatological probability of the event, which means, for every month 33% of the ensemble members are above the two-third percentile (dashed orange line in Figure 3.4) and 67% below, then the forecast system is perfectly reliable. However, it does not have any added value compared to just using the climatological values derived from observations. Thus, it is important that a prediction system is able to produce non-climatological forecast probabilities.

In Figure 3.4 we consider the event of "TWS above the two-third percentile" (dashed orange line), and focus on high model probabilities (90 - 100%). A reliable forecast system is shown in Figure 3.4a, where in about 90% of the months with a model probability of 90 - 100%, the observations (orange line) lie above the two-third percentile (gray shaded areas vs. orange shaded areas). A lack of reliability can originate from a lack of ensemble spread, making the forecast overconfident, or from an excess of ensemble spread, making the forecast underconfident. An overconfident forecast system (Figure 3.4b) exhibits a small spread, which means that the ensemble member consensus is relatively high, so that high probabilities for an event are more frequent than the observed probabilities are. In Figure 3.4b more occasions with model probability of 90 -100% (gray shaded areas) occur compared to Figure 3.4a, but only in about 70% of these occasions also the observations agree with this (orange shaded areas). Analogously, a large ensemble spread (Figure 3.4c) makes high probabilities for an outcome rare, and thus, the modeled probabilities are smaller than the observed. In Figure 3.4c less occasions with model probability of 90 -100% (gray shaded areas) occur compared to Figure 3.4a, but all of these are in agreement with the observations (orange shaded areas).

The reliability of a forecast system is often tested by comparing the mean ensemble spread to the RMSD between the ensemble mean and the observations (Palmer et al., 2006). The idea is that given a set of forecasts and an observational record, they should not be distinguishable by their variability. The RMSD between the ensemble mean and a single ensemble member should (on average) be the same as for the observational time series. In Figure 3.4a the RMSD (of observations vs. ensemble mean) is equal to the mean ensemble spread, and if the observational time series would not be highlighted in orange, it could hardly be distinguished from the model runs. In contrast, in Figure 3.4b (and c) the RMSD is larger (smaller) than the mean spread, and the observational time series can be distinguished from the ensemble members by its larger (smaller) variability.

3.5.3. Forecast-year time series

In contrast to the outcome of climate projections, time series of decadal predictions (or hindcasts, if run for the past) are directly comparable to observational data sets because they are supposed to represent the real climate conditions on specific points in time. By comparing decadal predictions to observations, their predictive skill can be assessed. A frequently applied method to find out how many years into the future the initialization has an effect, i.e. how skillful the predictions are, is the comparison of so-called forecast-year time series to observations. This was also carried out in Paper No. 3. The idea is that—in case of a positive effect of initialization—the match between modeled and observed TWS should be best in the first year after initialization and then gradually decrease with increasing forecast year.

Therefore, the decadal time series of annual mean TWS anomalies are rearranged according to their different forecast years (Figure 3.5): Since the decadal simulations are initialized every year, forecast year 1 modeled TWS anomalies (blue dots in Figure 3.5) exist for each year between 1961 and 2010 (the time span of decadal experiments in CMIP5). By keeping only the first year anomalies from each decadal hindcast we obtain a time series consisting only of forecast year 1 anomalies (blue line in Figure 3.5). Analogously, a forecast year 2 time series (orange curve) is obtained from just keeping the year 2 anomalies (orange dots) from each decadal run. This results in ten time series (one for each forecast year) covering the time period 1970 - 2010 (1970 is the first year for which the 10th forecast year exists).

These forecast-year time series (blue, orange and yellow lines in Figure 3.5) can now be compared to an observational record, e.g. in terms of correlation and RMSD. Usually the correlation is expected to decrease with increasing forecast year (RMSD increases). For a reference, the correlation or RMSD values computed for the different forecast-year time series are compared to the correlation or RMSD values of an uninitialized projection, i.e. a time series from a long-term climate projection run, where the initialization in 1850 has virtually no influence any more. Only if the correlation (RMSD) of the observations with the forecast-year time series is higher (lower) than with the uninitialized projection, the prediction has skill.

3.6. Terrestrial water storage in Earth System Models

Coupled climate models provide global fields of the status for a large number of variables for all major components of the Earth system for each time step (e.g. each month). In CMIP5 and CMIP6 the land water storage-related variables, which are suitable to be compared to TWS satellite observations are ”total soil moisture content” (*mrso*) and ”surface snow amount” (*snw*). In all three papers of this thesis we used the sum of these two variables to approximate modeled terrestrial water storage (mTWS). For specific purposes we also analyzed *mrso* and *snw* separately.

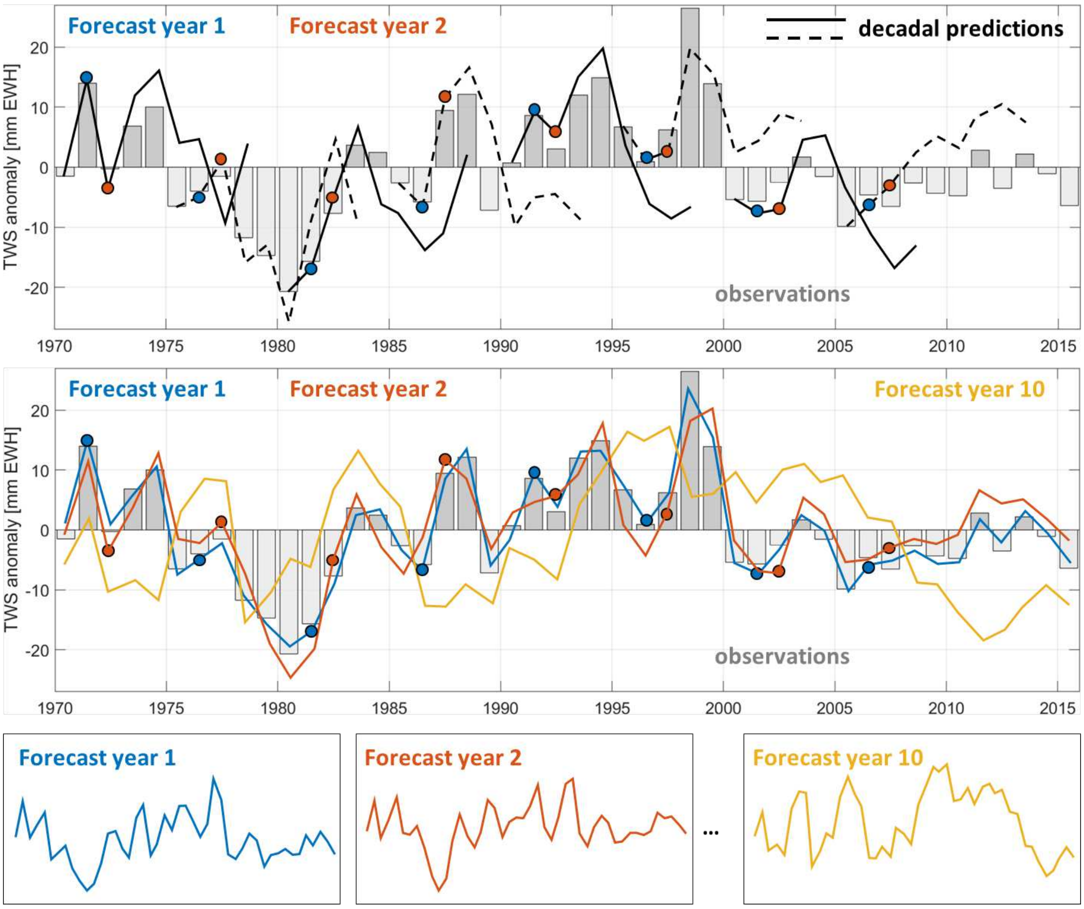


Figure 3.5.: Schematic illustration of the building of forecast-year time series. Upper panel: Observations (gray bars) and decadal predictions (black solid and dashed lines) initialized every 5 years (for graphical representation). The first forecast year anomaly for each decadal prediction is depicted with a blue dot, the second with an orange dot. Middle and lower panels: All forecast year 1 anomalies (blue dots) are connected and build the forecast year 1 time series (blue line), which is relatively close to the observations. Please note that a forecast year 1 anomaly for each year exists (due to yearly initialization), but only every fifths is marked with a blue dot (for graphical representation). Analogously, all forecast year 2 anomalies (orange dots) are connected to forecast year 2 time series (orange line), which is a bit less close to the observations than the forecast year 1 time series. All forecast year 10 anomalies build the forecast year 10 time series (yellow line), which is far off the observations.

The modeled TWS does not explicitly contain groundwater modeling, surface water variations, or mass loss from glaciers and ice caps. Furthermore, it does not include human interventions into the water cycle, such as groundwater depletion or dam building. Therefore, GRACE-derived TWS and mTWS do not represent the same physical entity at every location. In the three papers of this thesis we extensively discuss these discrepancies and their influence on the results of the respective study. We furthermore identify regions where the differences may be particularly large, and where the results therefore have to be handled with care or should even be masked out.

4. Research Objectives

In this chapter, the main research objectives of this thesis are outlined. Section 4.1 can be seen as an overarching research objective, dealing with the general possibilities of the comparison of model output and GRACE observations. It is divided into sub-aspects and particularized in the following sections 4.2, 4.3, and 4.4, which correspond to the subjects of the three papers of this thesis.

4.1. Comparability of climate model output and GRACE/-FO observations

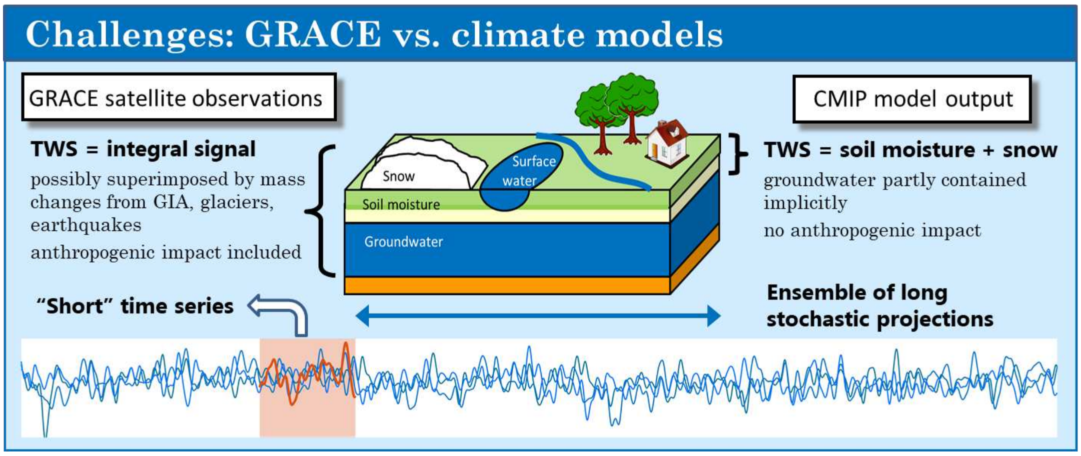
In Chapter 3 several challenges are pointed out that arise when comparing climate model output to space gravimetric observations (schematic summary in Figure 4.1). Generally, these challenges depend on whether climate projections or climate predictions are examined. Climate projections reproduce the climatic conditions in a statistical manner only (cf. Section 3.4) and therefore cannot directly be compared to observations on time series level. The two aspects discussed in Section 3.4, the internal variability and the dampening through averaging, prohibit the comparison. In contrast, decadal climate predictions can be directly evaluated with observations by means of skill scores based on measures like the correlation coefficient or the RMSD. This leads to two overarching research questions for this thesis:

1. How can we use space gravimetric observations to evaluate the performance of climate model projections devoid of the direct comparison of time series?
2. Are space gravimetric observations suitable to assess the skill of decadal land water storage predictions from climate models?

The first question is dealt with in Papers No. 1 and 2. They highlight specific aspects of the comparison of observations and climate models, which are formulated in more detail in Sections 4.2 and 4.3. The second question is considered in Paper No. 3, with specific research questions provided in Section 4.4.

4.2. Long-term trends in terrestrial water storage

Long-term drying and wetting trends have a direct effect on the availability of freshwater resources and therefore are of great societal importance. However, TWS signals are a composition of long-term, annual, interannual and sub-seasonal variations, and over relatively short time spans (20 to 30 years) linear wetting and drying trends are likely to be obscured by interannual variations (Figure 4.2). In this context, the following research questions were derived, which are investigated in Paper No. 1:



Challenges: GRACE vs. climate models

GRACE satellite observations

TWS = integral signal

possibly superimposed by mass changes from GIA, glaciers, earthquakes

anthropogenic impact included

"Short" time series

CMIP model output

TWS = soil moisture + snow

groundwater partly contained implicitly

no anthropogenic impact

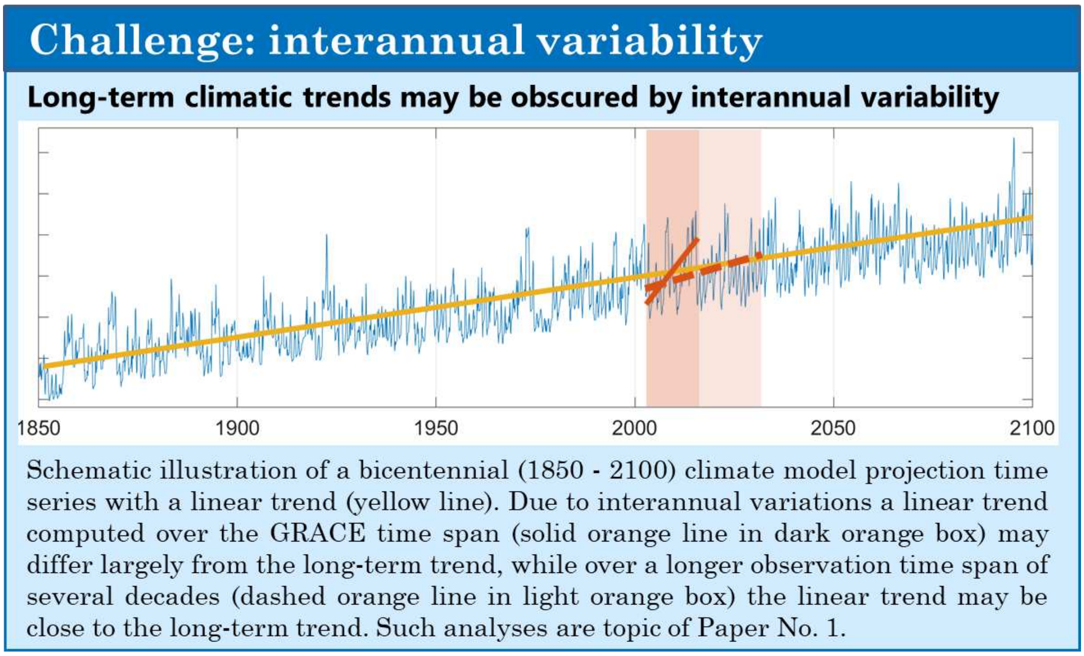
Ensemble of long stochastic projections

Figure 4.1.: Challenges for comparing climate model output and GRACE observations resulting from (1) conceptual discrepancies of TWS in models and observations, (2) the stochastic character of climate projections and (3) the short observational record.

1. How well do climate model projections agree on long-term TWS trends in terms of magnitude and spatial distribution? As some climate models in the CMIP5 data base are only sub-versions of each other or share common components, we first have to investigate: How many and which models do actually provide independent output for the analysis?
2. What degree of agreement between long-term (bicentennial) and short-term (14 years) TWS trends can be expected in the presence of overlying interannual variations, globally and regionally? Do GRACE-derived TWS trends meet these expectations? How long do we have to observe to reliably distinguish between interannual variations and long-term trends?
3. Can we confirm climate-related wetting or drying conditions by means of GRACE in regions of high model consensus?

4.3. Emerging changes in terrestrial water storage variability

Climate change will affect terrestrial water storage during the next decades by impacting not only long-term linear trends but also the seasonal cycle and interannual variations. To date, observational time series are too short to find changes in the variability of TWS, e.g. increasing/decreasing amplitude of the seasonal cycle or increasing/decreasing magnitude of interannual variations. However, an intensification of the seasonal hydroclimate cycle in some regions was already found in precipitation records and attributed to amplifications in the atmospheric moisture budget (Liang et al., 2020; Salerno et al., 2019). The seasonal water cycle was also found to be shifting over time regionally, for example due to a climate-driven later onset of the rainy season (Luković et al., 2021). Such alterations of the terrestrial water cycle are very likely to continue or even intensify in the coming decades, and long-term projections from ESMs can provide information about the future development of the variability. The following research questions (illustrated in Figure 4.3) are related to changes in TWS variability, and addressed in Paper No. 2:



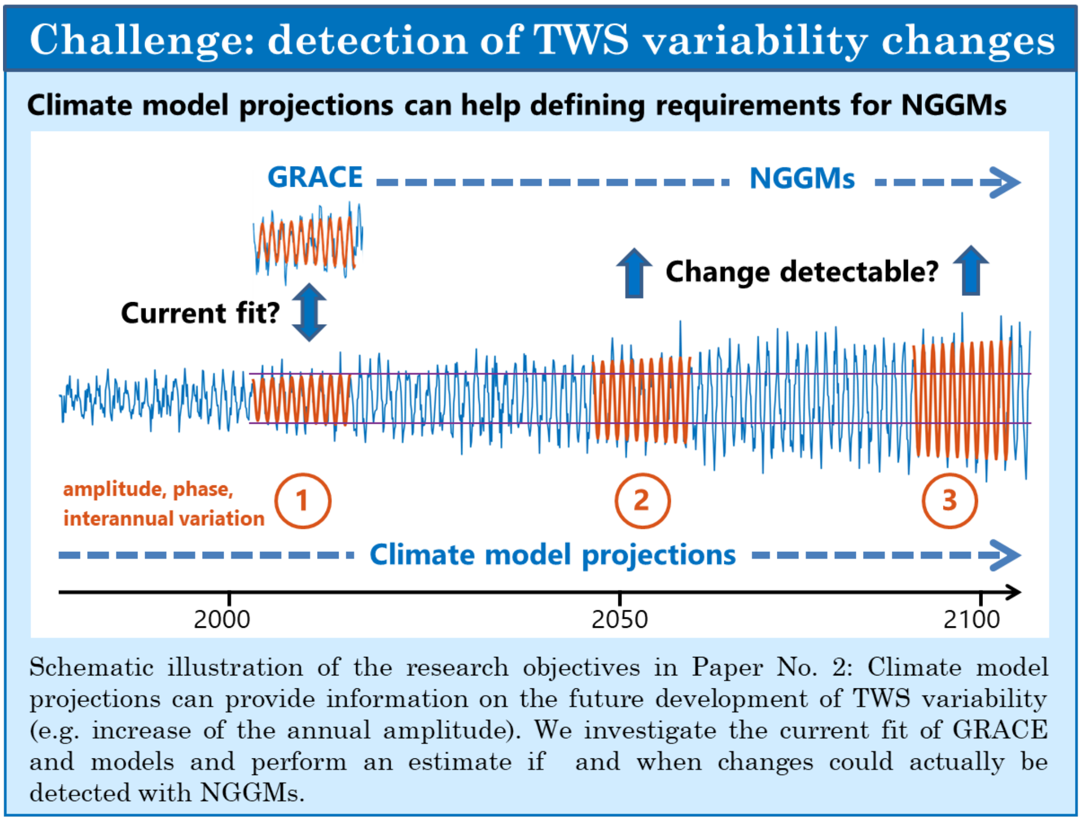
Challenge: interannual variability

Long-term climatic trends may be obscured by interannual variability

Schematic illustration of a bicentennial (1850 - 2100) climate model projection time series with a linear trend (yellow line). Due to interannual variations a linear trend computed over the GRACE time span (solid orange line in dark orange box) may differ largely from the long-term trend, while over a longer observation time span of several decades (dashed orange line in light orange box) the linear trend may be close to the long-term trend. Such analyses are topic of Paper No. 1.

Figure 4.2.: Schematic illustration of interannual variations in TWS time series obscuring long-term trends.

1. Model-derived results about changes in TWS variability have limited value if the climate models are not reliable. Thus it is important to address the question: How do modeled and observed TWS agree in terms of annual cycle and interannual variations in the GRACE time span? A reasonable comparison requires a precedent investigation: How can we derive a "best estimate" of the annual cycle and interannual variations from a multi-model ensemble?
2. Which changes in TWS variability are to be expected in the coming decades according to ESMs and how concordant are the ESMs regarding these changes?
3. Model-derived estimates on the magnitude and spatial distribution of changes in TWS variability are needed to answer the question: How long do we have to observe to be able to identify such changes in space gravimetric data records? To which extent are the expected changes detectable with a GRACE-like satellite mission or with a Next Generation Gravity Mission?
4. Long time series of TWS as provided by model projections are also needed for long-term NGGM simulation studies that demonstrate the value of satellite gravimetry for monitoring climate signals in TWS. However, using the multi-model mean as input time series is not appropriate because interannual variations are largely smoothed out during averaging. This leads to the question: How can we select a representative model run from the ensemble that can serve as input for NGGM simulation studies?



Challenge: detection of TWS variability changes

Climate model projections can help defining requirements for NGGMs

Schematic illustration of the research objectives in Paper No. 2: Climate model projections can provide information on the future development of TWS variability (e.g. increase of the annual amplitude). We investigate the current fit of GRACE and models and perform an estimate if and when changes could actually be detected with NGGMs.

Figure 4.3.: Schematic illustration of TWS variability (depicted as annual cycle in orange) in climate model projections. It may (1) fit to current observations, and (2), (3) change over the projection time span, which may be detectable with future gravity missions.

4.4. Decadal predictability of terrestrial water storage

Information about the evolution of TWS a few years into the future would be relevant for agricultural and water management decisions. Therefore, the skill of decadal predictions has to be assessed. Decadal predictions do not depend on emission scenarios but provide unconditional forecasts. Therefore, their output should in principle be directly comparable to observations. Within this context, the following research questions are formulated, which are addressed in Paper No. 3:

1. How many years into the future are decadal predictions superior to uninitialized projections or trivial forecasts regarding the predictive skill? How large is the predictive skill on global, regional and grid cell level?
2. Are decadal predictions of TWS reliable?
3. Are there indications that CMIP6 decadal hindcasts have an improved predictive skill compared to those from CMIP5?

5. Results and Conclusions

In this chapter compact answers to the research questions formulated in Chapter 4 are given. They provide an overview of the main results of this thesis. A detailed discussion of the research objectives and results can be found in the three papers that are reprinted in Appendix A.

5.1. Comparability of climate model output and GRACE/-FO observations

Q1: How can we use space gravimetric observations to evaluate the performance of climate model projections devoid of the direct comparison of time series?

While the one-to-one comparison of model and observational time series is not possible for long-term projections from coupled ESMs, the derivation of higher order metrics described by deterministic or stochastic parameters enables an assessment of model abilities to represent the signal variability by means of observations. These higher order metrics require, e.g., a decomposition of the time series into different signal components, the derivation of extreme value statistics, or the computation of more abstract measures, such as soil moisture memory (illustrated in Figure 5.1):

Variability measures
The decomposition of the time series into linear trend, seasonal cycle, interannual and sub-seasonal signal can be utilized to compare the variability in the different signal components. As the linear trend and seasonal cycle in ESMs are generally driven by the prescribed external forcing, they can be characterized by estimating deterministic parameters for trend, annual amplitude and phase. In contrast, the interannual and sub-seasonal signals in ESM projections are governed by the internal model variability. Therefore, such signals can only be compared based on the magnitude of their variability, e.g., by computing the temporal root mean square (RMS) over a certain time span. For individual model runs, these quantities can then be compared to the observations by analyzing the spatial pattern or the distribution of values. However, when analyzing a multi-model ensemble, simple averaging to a multi-model mean is not applicable, as the ensemble variability is largely smoothed out. Strategies have to be found to compute "best estimates" of the measures while maintaining the ensemble's variability. Furthermore, over short time periods, i.e., less than 30 years, the distinction between long-term linear trends and superimposed interannual variations is difficult. Therefore, the influence of interannual variability on different time scales has to be thoroughly assessed before conclusions on the fit between modeled and observed trends can be drawn.

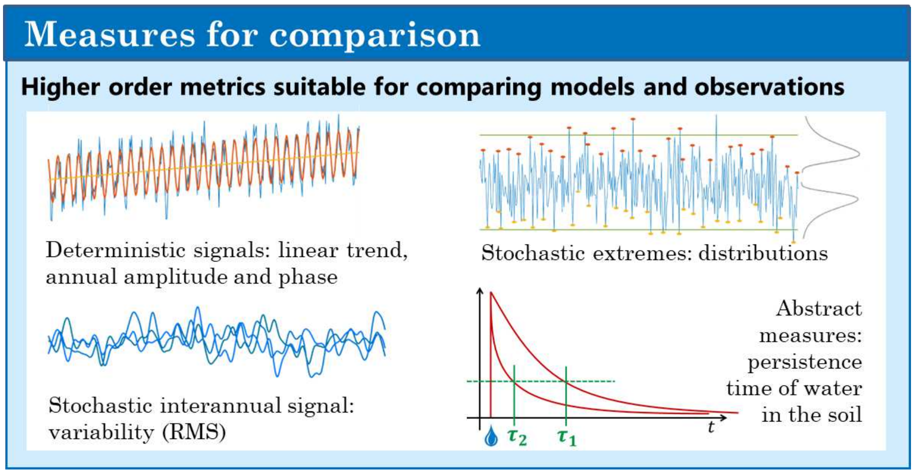

Measures for comparison
Higher order metrics suitable for comparing models and observations
Deterministic signals: linear trend, annual amplitude and phase
Stochastic extremes: distributions
Stochastic interannual signal: variability (RMS)
Abstract measures: persistence time of water in the soil
τ_2
τ_1
t

Figure 5.1.: Measures to compare climate projections and GRACE.

Extreme events

The ability of models to reproduce the recurrence frequency of extreme events is important in order to gauge their future impact on our society. Extreme events in land water storage manifest themselves as droughts and floods, which both can have severe consequences for (human) life, society, and economic prosperity in affected regions. The relative magnitude and recurrence frequency of extreme TWS events are two parameters that can be compared between models and GRACE/GRACE-FO observations to assess the current ability of the models to simulate extreme events. Again, the analysis has to be performed on the level of individual model runs, since the time series are stochastic and therefore averaging would smooth out extreme values. The question if, how, and where extreme events in TWS are intensifying due to climate change cannot yet be answered from the relatively short GRACE/GRACE-FO record, but long-term climate projections can help to quantify the time span that would be needed to observe such changes.

Abstract measures

A different option to use observations for the evaluation of climate models is to derive abstract measures that describe internal model parameters. An example of such a parameter is the soil moisture memory (SMM), describing the average residence time of water in the soil. This parameter can be computed by filtering and correlating precipitation data to TWS data, because past precipitation events are reflected in land water storage with a certain time lag, which is related to SMM (Humphrey et al., 2016). SMM can be computed from climate models by using the respective output variables precipitation, soil moisture, and snow from each model. It can be compared to observed SMM derived from precipitation data sets and TWS from GRACE/GRACE-FO. In doing so, the general representation of land water storage and soil moisture processes in climate models can be validated, and potential deficits can be revealed.

Within this thesis (in Papers No. 1 and 2), the comparison of models and observations is realized by the investigation of different variability measures. Stochastic events like the occurrence of extreme values or more abstract measures remain topic of future work.

Q2: Are space gravimetric observations suitable to assess the skill of decadal land water storage predictions from climate models?

In principle, they are. However, there are regional differences in the quality of the skill assessment. This question is answered in more detail in Paper No. 3 (Section 5.4).

5.2. Long-term trends in terrestrial water storage

Q1: How well do climate model projections agree on long-term TWS trends in terms of magnitude and spatial distribution? How many and which models do actually provide independent output for the analysis?

In the CMIP5 data base, 34 models provide modeled TWS (the sum of total soil moisture content and surface snow amount, abbreviated to mTWS) for the historical and RCP8.5 experiments. Independent models can be identified by comparing the similarity of mTWS trend maps for all 34 models. The correlation coefficient calculated from the vectorized maps serves as the measure of similarity. Models with a particularly high pattern correlation (>75%) between their mTWS trend maps are considered to be not independent. By excluding all but one from the highly correlated models, 21 of 34 models remain for further analysis.

Generally, the long-term (bicentennial) trend maps of the remaining 21 models exhibit a low correlation of 10% on average, demonstrating a large inhomogeneity among CMIP5 models regarding mTWS trends. Therefore, we focused on the more general investigation of wetting or drying trends regardless of their magnitude. We derived a consensus map for the 21 models from counting for each grid cell the number of models agreeing on a negative or positive mTWS trend. We identified several regions of high model consensus (39% of the global land area) where at least 71% of the models (15 of 21) agree on the sign of the trend. In the majority of the high consensus regions, models see a drying (77%) rather than a wetting (23%) trend. A reason for underrepresented wetting trends in CMIP5 models might be that the models have a limited ability to capture anomalously high water storage, since excess water is allocated to runoff into the ocean. Furthermore, an underestimation of soil moisture memory might prevent high accumulations of water.

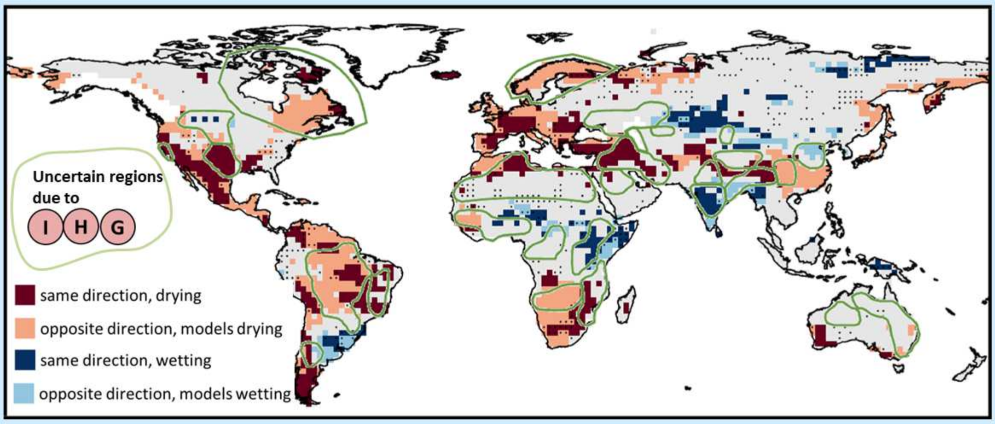
Q2: What degree of agreement between long-term (bicentennial) and short-term (14 years) TWS trends can be expected in the presence of overlying interannual variations, globally and regionally? Do GRACE-derived TWS trends meet these expectations? How long do we have to observe to reliably distinguish between interannual variations and long-term trends?

The degree of agreement between long-term and short-term trends was assessed in two numerical model investigations. The first investigation focused on the global scale. mTWS trend maps for differently long time spans (from 14 to 100 years) were computed and compared to the bicentennial trend map. From this model investigation, we concluded that 30 years would be the minimum time span to obtain a global pattern of TWS trends that can be mostly attributed to long-term climate trends rather than interannual climate variability. From 14 years of data we computed a global pattern correlation of (only) 23% with the bicentennial mTWS trend, which is similar to the actual correlation found for the observed 14-year GRACE TWS trend map (20%). The slightly higher value for the model data can be explained by the fact that the trend maps for the model investigation were not obtained from independent data, therefore representing somewhat idealized results.

While the first model investigation revealed what to expect from the similarity of the global spatial patterns, it did not provide the likelihood for a local mTWS trend computed from a certain time span to actually match the bicentennial trend in that grid cell. This was assessed in a second numerical model investigation: According to the models the probability that in a grid cell a 14-year trend is in agreement with the long-term trend is on average 53%, which is slightly better than random chance. While in the global pattern long-term trends emerge after about 30 years (first model investigation), for individual grid cells there is still a chance of 27% even after a century that the 100-year trend does not match the bicentennial trend. This indicates that there is regional natural variability in the models even over long time periods of a century and more. According to Laepple & Huybers (2014) the reason can be sea surface temperature varying on these time scales.

Highlight of Paper No. 1: Long-term trends

Agreement/disagreement of trends from GRACE and models



Regions of high model consensus regarding the direction of bicentennial mTWS trends, distinguished into regions where the TWS trend from GRACE has the same (dark red/blue) or opposite (light red/blue) direction. Regions affected by interannual variability (I), human impact (H) or glacial isostatic adjustment (G) are outlined by green polygons. Stippling indicates regions with non-significant GRACE trends.

Figure 5.2.: Highlight from Paper No. 1 (modified from Figures 7 and 8 of the paper).

Q4: Can we confirm climate-related wetting or drying conditions by means of GRACE in regions of high model consensus?

Comparing the sign of observational 14-year GRACE trends to the bicentennial mTWS trends on grid cell level, we found a 50% to 50% proportion for agreement and disagreement. The reason for the lower percentage of agreement compared to the model investigation (53%) is likely the idealized conditions in the model experiment. It can also result from an underestimation of interannual variability in the models. However, the analysis of the existing compliance and noncompliance regions especially in regions of high model consensus can give valuable information on potential climate-related wetting and drying as well as indications for possible model shortcomings.

By accessing a study by Rodell et al. (2018), regions where long-term wetting or drying trends may be dominated by interannual variability or human impacts were outlined. In these regions, which make up 36% of the high consensus area, it remains undetermined at this time if and how much they are affected by climate change. In the other 64% of the high consensus area the climate signal was concluded to dominate the TWS trends. Within these 64% the proportion of agreement vs. disagreement areas between the bicentennial mTWS trend and the GRACE trend is 49% vs. 51%. From this we concluded that in half of the area marked as climate-related wetting or drying, model deficits or unidentified interannual variability in observations cause disagreement. The other half we regarded as hot spot regions of drying and wetting that can already be attributed to climate change with the help of GRACE. Three large and spatially coherent areas could be identified: drying conditions in southwestern United States/Mexico and around the Mediterranean Sea, and wetting conditions in parts of Central Asia (Figure 5.2).

5.3. Emerging changes in terrestrial water storage variability

Q1: How do modeled and observed TWS agree in terms of annual cycle and interannual variations in the GRACE time span? How can we derive a "best estimate" of the annual cycle and interannual variations from a multi-model ensemble?

Before modeled and GRACE-derived results can be compared, a multi-model "best estimate" has to be derived from the ensemble of available climate model runs. Due to the stochastic character of climate models, the computation of a multi-model median (MMMed) time series from all ensemble members results in a reduced interannual and sub-seasonal variability. Thus, to maintain the complete variability of the model ensemble, we decompose the individual ensemble member time series separately into trend, annual cycle, interannual and sub-seasonal variations. Only afterward the MMMed grid over the annual amplitude and phase as well as the RMS of the interannual signal is computed. However, as the global patterns differ for individual ensemble members, the median smooths out extreme values in each grid cell. To counteract this, we apply a rescaling of the range of values in the MMMed based on the empirical cumulative density functions (ECDF) of all ensemble members.

The scaled MMMed maps are regarded as "best estimates" for the annual cycle and interannual signal from the model ensemble, and are compared to the respective measures derived from GRACE/GRACE-FO observations. For the annual amplitude the global patterns are similar, with an underestimation by the models in the equatorial climate zone, and an overestimation at polar latitudes. For the phase of the annual cycle we found that the models precede the GRACE observations in about two thirds of the land area by about half a month on average. For the RMS of the interannual signal the similarities with observations are not as strong as for the annual cycle, and the intermodel spread is larger. Land areas where models underestimate the interannual signal w.r.t. GRACE (60%) exceed areas of overestimation.

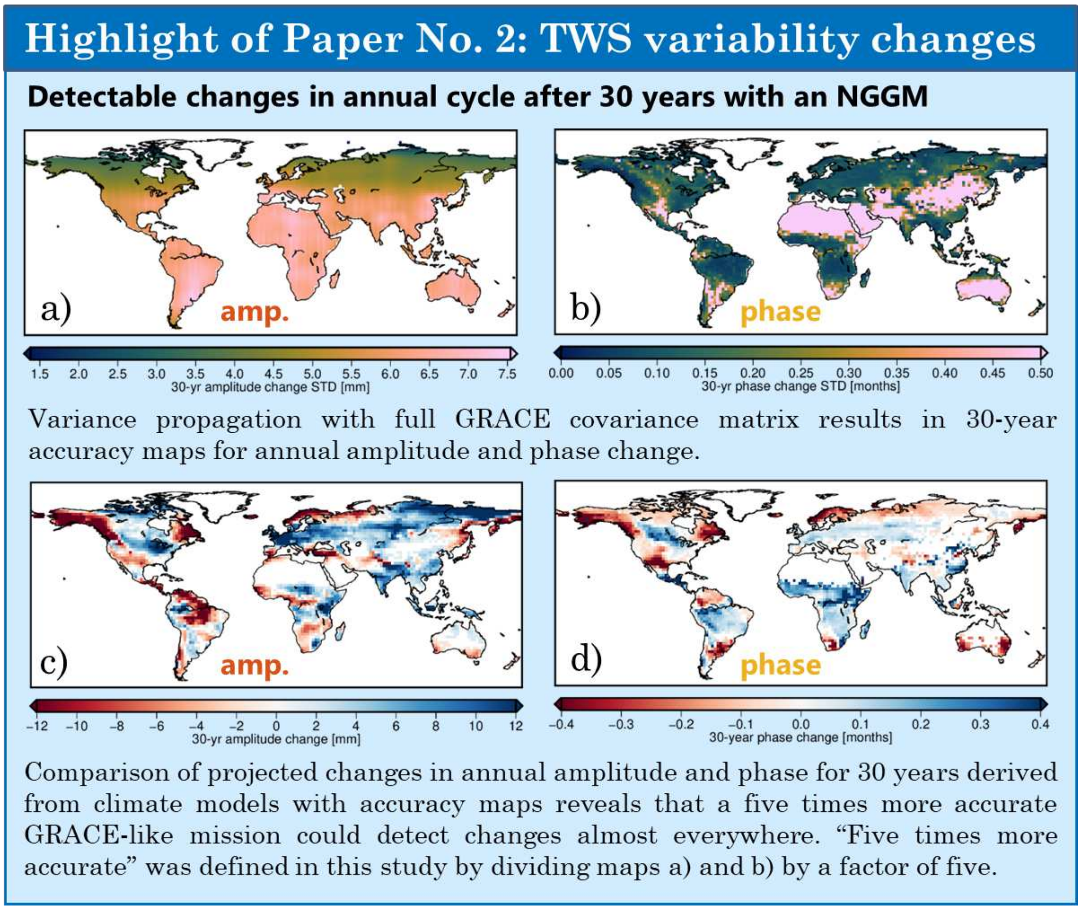
As GRACE observations are quite accurate and reliable for the measures and the spatial scales considered in Paper No. 2, discrepancies between modeled and observed variability can be mainly attributed to shortcomings in the models and missing water storage compartments. For example, the water holding capacity is supposedly bound-limited in models, leading to an overly strong runoff of excess water and subsequently smaller annual amplitudes. In addition, soil moisture memory is often too short in models especially in equatorial regions, preventing the accumulation of water to its real storage extent. The overestimation of the amplitude by the models in the north might be related to overestimated snow storage in winter and evapotranspiration in summer, thereby simulating an overall increased annual amplitude (Scanlon et al., 2019). The mainly positive phase shift between observations and models might be related to missing groundwater processes in CMIP6 models, causing an underestimation of the water residence time in the soil, hence less time for storage accumulation and consequently an earlier saturation of the maximum storage.

Q2: Which changes in TWS variability are to be expected in the coming decades according to ESMs and how concordant are the ESMs regarding these changes?

According to the CMIP6 models, changes in the annual amplitude of up to 27 mm per decade are to be expected regionally. In many regions (45% of the global land area) more than three-quarter of the models show the same direction of the amplitude change, which is positive in the majority of the land area (56%). Increases in TWS amplitude may be due to increased seasonal precipitation in a warmer climate (Chou & Lan, 2012) enhancing water storage in wet periods. The resulting amplitude increase may be amplified by increased evapotranspiration during summer, reducing water storage in dry periods. Decreasing amplitudes in polar regions may be related to generally rising temperatures and reduced snow accumulation. The models are less concordant about phase shifts of the annual cycle. In 37% of the land area three-quarter or more of the models agree on the direction of the shift. A particularly strong phase shift was found for equatorial regions, where until 2100 in 75% of the area the maximum of the annual cycle is projected to be reached on average over two weeks later. The reason is probably a later onset of the rainy season, which originates from changes in the atmospheric regimes. For example, in Africa, a position shift in the tropical rain belt and increasing strength of the Saharan heat low is projected (Dunning et al., 2018). The model consensus on changes of the interannual signal is still smaller than for the phase shifts; only in 23% of the land area more than three-quarter of the models agree on the change direction. Furthermore, the change pattern is more patchy than for the annual cycle with a tendency to an increased interannual variability (54% of the land area exhibits positive changes).

Q3: How long do we have to observe to be able to identify such changes in space gravimetric data records? To which extent are the expected changes detectable with a GRACE-like satellite mission or with a Next Generation Gravity Mission?

As both, the accuracy of space gravimetric data and the magnitude of expected changes, vary spatially, no absolute globally valid time span for the observation length needed to detect changes in TWS variability can be derived. Instead, we investigate for a fixed observation period of 30 years, in which regions changes in the annual amplitude and phase would be detectable with a mission maintaining the current GRACE accuracy, and alternatively also with a mission realizing an accuracy of five times higher than GRACE (which would be a plausible assumption for a double-pair NGGM). The regions were identified by comparing the respective accuracy patterns, obtained from rigorous variance-covariance propagation, from a representative GRACE monthly solution to the projected changes identified from the CMIP6 models. Whereas with a GRACE-like accuracy only in 34% (amplitude) and 28% (phase) of the land area changes would be observable after 30 years, with a mission of five times higher accuracy, these values would increase to 75% and 66%, where the missing areas are almost exclusively highly arid regions with barely any TWS variability (Figure 5.3).


Highlight of Paper No. 2: TWS variability changes

Detectable changes in annual cycle after 30 years with an NGGM

Variance propagation with full GRACE covariance matrix results in 30-year accuracy maps for annual amplitude and phase change.

Comparison of projected changes in annual amplitude and phase for 30 years derived from climate models with accuracy maps reveals that a five times more accurate GRACE-like mission could detect changes almost everywhere. "Five times more accurate" was defined in this study by dividing maps a) and b) by a factor of five.

Figure 5.3.: Highlight from Paper No. 2 (modified from Figure 11 of the paper).

Q4: How can we select a representative model run from the ensemble that can serve as input for NGGM simulation studies?

To select a representative model run from the CMIP6 multi-model ensemble we considered for each model run (1) its similarity of current TWS variability to GRACE observations and (2) its similarity of projected changes to the MMMed change of the respective component, i.e., annual amplitude and phase, interannual variation. As measures of similarity we used the pattern correlation and the RMSD of the ECDFs. The model runs were sorted according to the value of correlation and RMSD for each component, and a ranking was assigned. By this, the most representative model run regarding its mean fit to observations and to the MMMed change could be identified. The ranking also revealed that no single model run is clearly superior to all others, but that several model runs exhibit a similar mean fit. The individual fit for different measures of similarity varies between the different model runs, so that no model ranks highest in all categories. However, as the model run should only represent a typical system behavior, a "best fit" in all categories is not necessarily a requirement for NGGM simulations.

5.4. Decadal predictability of terrestrial water storage

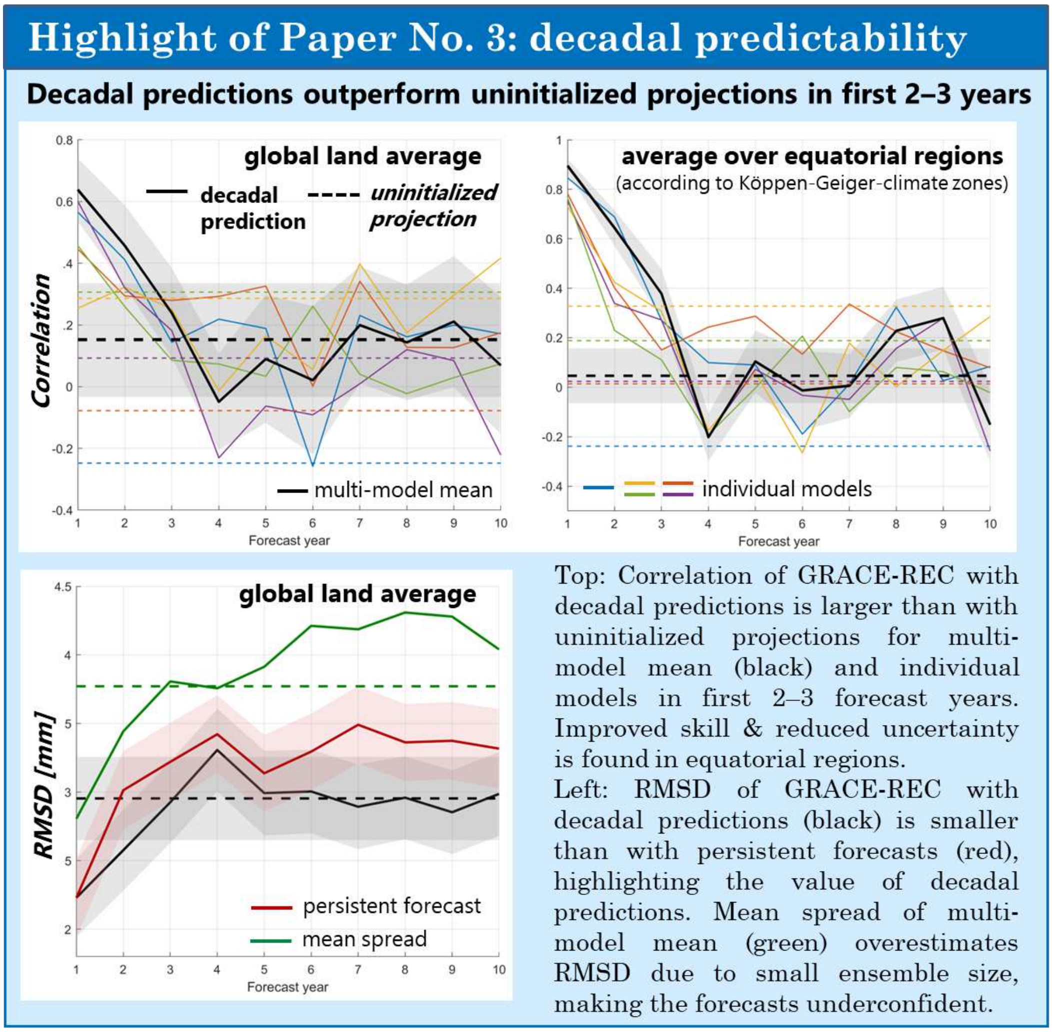
Q1: How many years into the future are decadal predictions superior to uninitialized projections or trivial forecasts regarding the predictive skill? How large is the predictive skill on global, regional and grid cell level?

To date, the temporal overlap between GRACE observations and decadal CMIP5 hindcasts is still too short for a robust comparison. Thus, to assess the predictive skill of land water storage in decadal predictions we resort to a century-long reconstruction of GRACE TWS anomalies (GRACE-REC by Humphrey & Gudmundsson (2019), cf. Section 2.3). Based on GRACE-REC as the observational reference the skill was evaluated with respect to different yearly forecast horizons. Thus, we created forecast-year time series (cf. Section 3.5) from the yearly-initialized hindcasts (abbreviated to Init) which were available from five CMIP5 models. Then we calculated the correlation and the RMSD w.r.t. GRACE-REC for the global land average. As a reference for the skill assessment, we used the uninitialized (abbreviated to Hist) experiments (historical concatenated with RCP4.5 simulations) from the respective models.

We found that for the global land average the Init simulations outperform the Hist runs for the first three forecast years (Figure 5.4), while the superiority in the third forecast year is not very distinct. The Init correlations for the first three forecast years are also higher than those obtained from a persistent forecast, where the state of the first forecast year is maintained over the complete forecast period. This underlines the added value of decadal predictions with respect to trivial forecasts.

In the regional analysis we repeated the skill assessment for time series averaged over different climate zones. While in arid, temperate, and polar regions the results for the Init simulations are degraded in comparison to the global analysis, in the equatorial climate zone much higher correlations and smaller RMSDs were found, even for the third forecast year. Furthermore, we computed global 2° maps for the correlation between

GRACE-REC and Init/Hist time series. From these, a general success of the initialization in forecast year 1 was identified, but a general regional prediction skill for TWS for lead times longer than 2 years was not found in CMIP5.



Highlight of Paper No. 3: decadal predictability

Decadal predictions outperform uninitialized projections in first 2–3 years

Top: Correlation of GRACE-REC with decadal predictions is larger than with uninitialized projections for multi-model mean (black) and individual models in first 2–3 forecast years. Improved skill & reduced uncertainty is found in equatorial regions.
Left: RMSD of GRACE-REC with decadal predictions (black) is smaller than with persistent forecasts (red), highlighting the value of decadal predictions. Mean spread of multi-model mean (green) overestimates RMSD due to small ensemble size, making the forecasts underconfident.

Figure 5.4.: Highlight from Paper No. 3 (modified from Figures 2, 3, and 4 of the paper).

Q2: Are decadal predictions of TWS reliable?

To assess the reliability of decadal mTWS predictions we compared the ensemble spread to the RMSD between Init and GRACE-REC anomalies. We found that the model spread generally reflects the rise of the RMSD with increasing forecast year, but overestimates it by a factor of about 1.3, indicating that the models are underconfident for that particular metric. This means that due to the relatively large ensemble spread (probably originating from the small ensemble size), the models forecast lower probabilities for specific TWS anomalies than actually observed.

Q3: Are there indications that CMIP6 decadal hindcasts have an improved predictive skill compared to those from CMIP5?

At the time of writing Paper No. 3, decadal hindcasts for three CMIP6 models were available, for which the forecast skill was assessed analogously to the CMIP5 models. From only these three models, the general level of prediction skill of the global average for the first three forecast years was found to be similar for CMIP5 and CMIP6. An improved reliability of CMIP6 in the early forecast years might be indicated by the smaller mean ensemble spread compared to CMIP5. When looking at individual models, we noticed a clear improvement from CMIP5 to CMIP6 for the model MPI-ESM only, which might be due to the fact that in MPI-ESM a new five-layer soil-hydrology scheme was used for CMIP6. This indicates a positive impact of enhanced hydrology schemes on the predictive skills of decadal simulations regarding TWS.

6. Summary and Outlook

6.1. Summary

Global coupled Earth System Models (ESMs) are an important tool for predicting future climate conditions of our planet. The evaluation of ESMs against observations is crucial for assessing their quality and reliability. In this thesis, land water storage-related variables from ESMs taking part in the Coupled Model Intercomparison Project Phases 5 and 6 (CMIP5 and CMIP6) were compared to observed terrestrial water storage (TWS) from the Gravity Recovery And Climate Experiment (GRACE) and its Follow-On (GRACE-FO) satellite missions. Several challenges were addressed during this comparison:

As GRACE and GRACE-FO measure integral mass changes regardless of their origin, observed TWS changes may be superimposed by non-hydrological mass changes, e.g., from glaciers, earthquakes, or glacial isostatic adjustment (GIA). Furthermore, GRACE-derived TWS is a composite of soil moisture, snow, surface water, and groundwater, while coupled ESMs only provide total soil moisture content and surface snow amount as land water storage-related variables. ESMs also do not include direct anthropogenic intervention into the water cycle. Therefore, in this thesis, locations of substantial mass variability due to anthropogenic groundwater abstraction, surface waters, glaciers, earthquakes, and GIA are identified using external data sets. These regions were either excluded from the analysis or marked as areas where results have to be interpreted with care.

Apart from the discrepancies between observed and modeled TWS, time series of observations and models are not necessarily directly comparable to each other. Depending on the forecast time span, climate model runs are distinguished between (long-term) climate projections and (decadal) climate predictions. The former often run for a century or longer and depend on a prescribed forcing scenario that dominates the forecast uncertainty (conditional forecast), whereas the latter run for typically 10 years and only depend on the initial state (unconditional forecast). This discrimination into projections and predictions led to two different approaches to compare GRACE/GRACE-FO data and climate model output. When evaluating long-term projections, higher order metrics as the linear trend, seasonal cycle and the RMS of interannual variations were analyzed instead of a direct comparison of time series. In contrast, decadal predictions were directly compared on time series level in terms of correlation and RMSD.

Long-term linear trends were investigated in detail in **Paper No. 1**. We found a large inhomogeneity among models regarding long-term (bicentennial) trends. A numerical model investigation revealed that the degree of agreement between bicentennial and short-term (14 years) trends is relatively low due to interannual variations obscuring climate-related signals. From the models we derived an observation time span of 30 years to reliably distinguish interannual from long-term trends. However, locally the influence of interannual variations can be much longer. A map of compliance or noncompliance of observations and models was derived by comparing 14 year-long GRACE trends and bicentennial model trends in areas where no particular indications of strong interannual variations or human impact are present and where models are highly concordant. This map revealed several distinct regions of potential climate-related wetting and drying (e.g., the Mediterranean, Southwestern United States, Central Asia), or regions with possible model shortcomings.

The annual cycle and interannual variations were the focus of **Paper No. 2**. From the multi-model ensemble best estimates for the annual amplitude, phase, and interannual RMS were derived by rescaling the multi-model median of each measure. GRACE and models fit reasonably well for the annual cycle and interannual RMS in terms of pattern correlation and relative differences. Regional clusters of under- or overestimation by the models with respect to observations are supposedly due to model shortcomings. They presumably originate, e.g., from limited water storage capacity, short soil moisture memory, deficiencies in land-atmosphere interactions, or missing groundwater storage. From the multi-model ensemble we derived the expected magnitude of changes in amplitude, phase, and interannual RMS until 2100. For the amplitude we found a high model agreement on the direction of the change in almost half of the land area with a regional dominance of increasing amplitudes. Phase shifts until the end of the century were projected to be quite substantial (up to several weeks) in both directions, while model agreement was slightly lower than for the amplitude. It was even lower for changes in the interannual variability, preventing a clear conclusion for this metric. Reasons for projected regional changes of TWS variability in a warmer climate may be increased evapotranspiration reducing water storage in summer, decreased snow accumulation, or changes in the atmospheric moisture transport, causing temporal shifts of the rainy season. We found that after 30 years, with a satellite mission exhibiting an accuracy that is five times better than the current GRACE accuracy, which is in reach for a Next Generation Gravity Mission (NGGM), projected changes in the annual cycle could be detected almost everywhere on the continents. Furthermore, to serve as input for NGGM simulation studies, we selected a specific model run that closely matched both GRACE observations and the multi-model median.

Paper No. 3 was dedicated to the decadal prediction skill of CMIP models. In principle, observations can be directly compared to decadal prediction time series. However, in this case, the overlap time span between decadal predictions and GRACE was too short for deriving robust results. Therefore, we resorted to a global reconstruction of TWS that is based on GRACE data (GRACE-REC) as a proxy for observed TWS. By means of GRACE-REC it could be shown that the decadal predictions outperform non-initialized projections for up to three forecast years, especially in equatorial regions. However, due to a large model spread, the reliability of the predictions is not very strong. A clear improvement in decadal prediction skill from CMIP5 to CMIP6 could not yet be proven in this study due to a small number of models available, but there were indications that an improved hydrology scheme has a positive impact on the prediction skill of the model.

Overall, this thesis was the first study to provide a comprehensive assessment of ways in which satellite gravimetric measurements can be utilized to evaluate coupled Earth System Models. It was shown that satellite-observed TWS with its unique integral character and global coverage has an excellent potential to validate model performance for land water storage-related variables in ESMs, and can also help to identify regions of model shortcomings and give hints at possible reasons for these. Vice versa, also climate model projections of TWS variability are of great value for advances of climate applications of satellite gravimetry, for example by enabling an estimation of necessary observation time spans or serving as input for NGGM performance studies.

6.2. Outlook

Current limits for the evaluation of climate models by means of GRACE observations are, amongst others, set by discrepancies of the water storage compartments contained in GRACE data and model output. While GRACE senses the integral signal of all mass changes regardless of their source, land water storage in ESMs is so far represented only by soil moisture and snow. To overcome this limitation, further effort must be put on the development of methods to separate different signals of the GRACE record. As such methods rely on external data obtained from other observation types or hydrological models, they benefit from advances made in these fields. Recently, a data set for correcting GRACE TWS for global lakes/reservoirs and earthquakes has been published (Deggim et al., 2021), and also techniques for the quantification of groundwater signals in GRACE observations are increasingly being explored (Frappart & Ramillien, 2018). Applying such corrections to the GRACE data to isolate the soil moisture and snow signal could help to make modeled and observed quantities more comparable.

In the context of Paper No. 1, the challenge of interannual variations masking long-term climate signals was highlighted. In that study locations suspected to be largely influenced by interannual variations were marked and excluded from further analysis. However, efforts to actually disentangle external forcing factors (likely dominated by human-induced global warming) and internal natural variability (i.e. interannual climate modes such as El Niño-Southern Oscillation, Pacific Decadal Oscillation, or Atlantic Meridional Oscillation) were already made, e.g. for the global mean sea level record (Moreira et al., 2021). Further exploring methods to quantify internal climate variability also for TWS (e.g., similar to Eicker et al., 2016) could help remove large parts of the interannual variations leading to a more rigorous isolation of long-term climatic trends in the GRACE record. This would improve the comparison of trends from even relatively short observational time series to long-term model trends. Furthermore, the assessment of the actual magnitude of model wetting and drying trends instead of restricting to the sign of the trend (as in Paper No. 1) would be feasible then.

The record of GRACE-FO is constantly growing and the importance of TWS in the climate system is increasingly being recognized, as expressed for example by the recent establishment of TWS as an ECV. Therefore, a continuation of satellite-based TWS measurements for the coming decades is likely. With the time series reaching the length of about 30 years, a more robust discrimination of long-term climate signals and internal climate variability would be possible, even directly from the observations. Increased spatial and temporal resolutions of NGGMs will pave the way for new climate applications of space gravimetric data and new possibilities of using them in climate model evaluation.

With the growing time series, it should also be possible to detect changes in the variability of TWS not only in model projections (as investigated in Paper No. 2), but also in the observations, potentially improving the validation of the model results. The identification of such changes could also benefit from advanced parametrizations, e.g., by not limiting seasonal variability to a sine/cosine signal, or by considering non-linear trends. In addition to variability measures, also probability density functions of TWS values in ESMs could be assessed by means of satellite gravity observations. Such assessments could also include spatio-temporal changes of the recurrence frequency and magnitude of extreme events, which is an important aspect in view of the severe consequences that floods and droughts often have. Moreover, the comparison of abstract measures like soil moisture memory, which can provide valuable information on land-atmosphere interactions in ESMs and potential shortcomings of modeling these processes, may contribute to model improvements.

Not only the signal separation of the observational record and methods for model evaluation will progress, but also new model generations will become available with altered representations of land water storage-related variables and land hydrology schemes. A possible implementation of additional components of the climate system, such as groundwater, surface water, or even anthropogenic impacts such as groundwater abstraction or dam building, into ESMs would be very beneficial for improved climate projections and would enhance the comparability to observations. Depending on the impact of new model parametrizations and the effect of a higher degree of model freedom and complexity, model spread and consensus will certainly change. Thus, evaluation by means of observations will continue to be important to judge new model developments.

First estimates on the detectability of changes in TWS variability with NGGMs were carried out in this thesis, but a simple variance propagation of uncertainty patterns from the current GRACE mission as was done here, is not sufficient. Full-fledged long-term performance studies for NGGMs are needed to derive realistic time spans and locations where changes in TWS variability could be detected with future missions. Criteria for the selection of a specific model run, which could serve as an input for such long-term NGGM simulations, were also proposed in this thesis. Thus, the next step would be the actual realization of an NGGM performance study based on such a model time series.

Advances in climate modeling and observation techniques will eventually bring ESMs and observations closer and closer together. The mutual benefit of using climate model output for the development of future gravity missions or even for signal separation of satellite-observed TWS, and using TWS observations for identifying and improving model shortcomings will continuously grow. In addition to TWS, also variables derived from other geodetic observation techniques have been used for the comparison to coupled ESMs. Examples for such variables are the dynamic sea surface height from satellite altimetry (Landerer et al., 2014), precipitable water vapor from GNSS (Roman et al., 2012), upper atmosphere temperature from GNSS radio occultation measurements (Kishore et al., 2016), or the hydrological excitation of polar motion from precise geodetic measurements of earth orientation parameters in combination with atmosphere and ocean models (Śliwińska et al., 2019). A consistent application of the approaches for model evaluation that were explored in this thesis to these geodetic data sets, could provide valuable insights into the representation of other variables and processes in ESMs. In a possible joint analysis of satellite gravimetry and other geodetic observations for a simultaneous validation of complementary ESM variables, synergy effects could be exploited, and the overall value of geodesy for climate model evaluation would raise even more.

References

Abich, K., Abramovici, A., Amparan, B., Baatzsch, A., Okihiro, B. B., Barr, D. C., ... Zimmermann, M. (2019). In-Orbit Performance of the GRACE Follow-on Laser Ranging Interferometer. *Physical Review Letters*, *123*(3), 031101. doi: 10.1103/PhysRevLett.123.031101

Baur, O., Kuhn, M., & Featherstone, W. E. (2009). GRACE-derived ice-mass variations over Greenland by accounting for leakage effects. *Journal of Geophysical Research: Solid Earth*, *114*(B6). doi: https://doi.org/10.1029/2008JB006239

Bellucci, A., Haarsma, R., Bellouin, N., Booth, B., Cagnazzo, C., Hurk, B. v. d., ... Weiss, M. (2015). Advancements in decadal climate predictability: The role of nonoceanic drivers. *Reviews of Geophysics*, *53*(2), 165–202. doi: https://doi.org/10.1002/2014RG000473

Bender, P., Wiese, D., & Nerem, R. S. (2008). A possible dual-grace mission with 90 degree and 63 degree inclination orbits. In *Proceedings of the 3rd international symposium on formation flying, missions and technologies* (p. 59-64).

Boer, G. J., Kharin, V. V., & Merryfield, W. J. (2013). Decadal predictability and forecast skill. *Climate Dynamics*, *41*(7), 1817–1833. doi: 10.1007/s00382-013-1705-0

Boer, G. J., Kharin, V. V., & Merryfield, W. J. (2019a). Differences in potential and actual skill in a decadal prediction experiment. *Climate Dynamics*, *52*(11), 6619–6631. doi: 10.1007/s00382-018-4533-4

Boer, G. J., Merryfield, W. J., & Kharin, V. V. (2019b). Relationships between potential, attainable, and actual skill in a decadal prediction experiment. *Climate Dynamics*, *52*(7), 4813–4831. doi: 10.1007/s00382-018-4417-7

Bryan, F. O., Danabasoglu, G., Nakashiki, N., Yoshida, Y., Kim, D.-H., Tsutsui, J., & Doney, S. C. (2006). Response of the North Atlantic Thermohaline Circulation and Ventilation to Increasing Carbon Dioxide in CCSM3. *Journal of Climate*, *19*(11), 2382–2397. doi: 10.1175/JCLI3757.1

Bunzel, F., Müller, W. A., Dobrynin, M., Fröhlich, K., Hagemann, S., Pohlmann, H., ... Baehr, J. (2018). Improved Seasonal Prediction of European Summer Temperatures With New Five-Layer Soil-Hydrology Scheme. *Geophysical Research Letters*, *45*(1), 346–353. doi: 10.1002/2017GL076204

Caron, L., Ivins, E. R., Larour, E., Adhikari, S., Nilsson, J., & Blewitt, G. (2018). GIA Model Statistics for GRACE Hydrology, Cryosphere, and Ocean Science. *Geophysical Research Letters*, *45*(5), 2203–2212. doi: 10.1002/2017GL076644

Cesana, G., & Waliser, D. E. (2016). Characterizing and understanding systematic biases in the vertical structure of clouds in CMIP5/CFMIP2 models. *Geophysical Research Letters*, *43*(19), 10,538–10,546. doi: https://doi.org/10.1002/2016GL070515

Chambers, D. P., Cazenave, A., Champollion, N., Dieng, H., Llovel, W., Forsberg, R., ... Wada, Y. (2017). Evaluation of the Global Mean Sea Level Budget between 1993 and 2014. *Surveys in Geophysics*, *38*(1), 309–327. doi: 10.1007/s10712-016-9381-3

Chen, J. L., Rodell, M., Wilson, C. R., & Famiglietti, J. S. (2005). Low degree spherical harmonic influences on Gravity Recovery and Climate Experiment (GRACE) water storage estimates. *Geophysical Research Letters*, *32*(14). doi: https://doi.org/10.1029/2005GL022964

Chen, J. L., Wilson, C. R., Li, J., & Zhang, Z. (2015). Reducing leakage error in GRACE-observed long-term ice mass change: a case study in West Antarctica. *Journal of Geodesy*, *89*(9), 925–940. doi: 10.1007/s00190-015-0824-2

Cheng, M., & Ries, J. (2017). The unexpected signal in GRACE estimates of C_{20}. *Journal of Geodesy*, *91*(8), 897–914. doi: 10.1007/s00190-016-0995-5

Cheng, M., Ries, J. C., & Tapley, B. D. (2011). Variations of the Earth's figure axis from satellite laser ranging and GRACE. *Journal of Geophysical Research: Solid Earth*, *116*(B1). doi: https://doi.org/10.1029/2010JB000850

Cheng, M., Tapley, B. D., & Ries, J. C. (2010). *Geocenter Variations from Analysis of SLR data* (Tech. Rep. No. 138). Marne-La-Vallée, France: Reference Frames for Applications in Geosciences, International Association of Geodesy Symposia.

Chou, C., & Lan, C.-W. (2012). Changes in the Annual Range of Precipitation under Global Warming. *Journal of Climate*, *25*(1), 222–235. doi: 10.1175/JCLI-D-11-00097.1

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., ... Wehner, M. (2013). Long-term Climate Change: Projections, Commitments and Irreversibility. In T. Stocker et al. (Eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 1029–1136). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. doi: 10.1017/CBO9781107415324.024

Committee on the Decadal Survey for Earth Science and Applications from Space, Space Studies Board, Division on Engineering and Physical Sciences, & National Academies of Sciences, Engineering, and Medicine. (2018). *Thriving on Our Changing Planet: A Decadal Strategy for Earth Observation from Space*. Washington, D.C.: National Academies Press. doi: 10.17226/24938

Cook, B. I., Mankin, J. S., Marvel, K., Williams, A. P., Smerdon, J. E., & Anchukaitis, K. J. (2020). Twenty-First Century Drought Projections in the CMIP6 Forcing Scenarios. *Earth's Future*, *8*(6), e2019EF001461. doi: https://doi.org/10.1029/2019EF001461

Corti, S., Weisheimer, A., Palmer, T. N., Doblas-Reyes, F. J., & Magnusson, L. (2012). Reliability of decadal predictions. *Geophysical Research Letters*, *39*(21), L21712. doi: 10.1029/2012GL053354

Dahle, C., Murböck, M., Flechtner, F., Dobslaw, H., Michalak, G., Neumayer, K. H., . . . Förste, C. (2019). The GFZ GRACE RL06 Monthly Gravity Field Time Series: Processing Details and Quality Assessment. *Remote Sensing*, *11*(18), 2116. doi: 10.3390/rs11182116

Daras, I., & Pail, R. (2017). Treatment of temporal aliasing effects in the context of next generation satellite gravimetry missions. *Journal of Geophysical Research: Solid Earth*, *122*(9), 7343–7362. doi: 10.1002/2017JB014250

Deggim, S., Eicker, A., Schawohl, L., Gerdener, H., Schulze, K., Engels, O., . . . Longuevergne, L. (2021). RECOG RL01: correcting GRACE total water storage estimates for global lakes/reservoirs and earthquakes. *Earth System Science Data*, *13*(5), 2227–2244. doi: 10.5194/essd-13-2227-2021

Doblas-Reyes, F. J., Andreu-Burillo, I., Chikamoto, Y., García-Serrano, J., Guemas, V., Kimoto, M., . . . van Oldenborgh, G. J. (2013). Initialized near-term regional climate change prediction. *Nature Communications*, *4*, 1715. doi: 10.1038/ncomms2704

Dunning, C. M., Black, E., & Allan, R. P. (2018). Later Wet Seasons with More Intense Rainfall over Africa under Future Climate Change. *Journal of Climate*, *31*(23), 9719–9738. doi: 10.1175/JCLI-D-18-0102.1

Döll, P., Schmied, H. M., Schuh, C., Portmann, F. T., & Eicker, A. (2014). Global-scale assessment of groundwater depletion and related groundwater abstractions: Combining hydrological modeling with information from well observations and GRACE satellites. *Water Resources Research*, *50*(7), 5698–5720. doi: 10.1002/2014WR015595

Eade, R., Smith, D., Scaife, A., Wallace, E., Dunstone, N., Hermanson, L., & Robinson, N. (2014). Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophysical Research Letters*, *41*(15), 2014GL061146. doi: 10.1002/2014GL061146

Eicker, A., Forootan, E., Springer, A., Longuevergne, L., & Kusche, J. (2016). Does GRACE see the terrestrial water cycle "intensifying"? *Journal of Geophysical Research: Atmospheres*, *121*(2), 2015JD023808. doi: 10.1002/2015JD023808

Eicker, A., Schumacher, M., Kusche, J., Döll, P., & Schmied, H. M. (2014). Calibration/-Data Assimilation Approach for Integrating GRACE Data into the WaterGAP Global Hydrology Model (WGHM) Using an Ensemble Kalman Filter: First Results. *Surveys in Geophysics*, *35*(6), 1285–1309. doi: 10.1007/s10712-014-9309-8

European Space Research and Technology Centre. (2020). *Next Generation Gravity Mission as a Mass-change And Geosciences International Constellation (MAGIC) - A joint ESA/NASA double-pair mission based on NASA's MCDO and ESA's NGGM studies.* ESA-EOPSM-FMCC-MRD-3785.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. doi: https://doi.org/10.5194/gmd-9-1937-2016

Fasullo, J. T., Lawrence, D. M., & Swenson, S. C. (2016). Are GRACE-era Terrestrial Water Trends Driven by Anthropogenic Climate Change? *Advances in Meteorology*. doi: 10.1155/2016/4830603

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S., Collins, W., . . . Rummukainen, M. (2013). Evaluation of Climate Models. In T. Stocker et al. (Eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 741–866). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. (Section: 9 Type: Book Section) doi: 10.1017/CBO9781107415324.020

Flechtner, F., Neumayer, K.-H., Dahle, C., Dobslaw, H., Fagiolini, E., Raimondo, J.-C., & Güntner, A. (2016). What Can be Expected from the GRACE-FO Laser Ranging Interferometer for Earth Science Applications? *Surveys in Geophysics*, *37*(2), 453–470. doi: 10.1007/s10712-015-9338-y

Frappart, F., & Ramillien, G. (2018). Monitoring Groundwater Storage Changes Using the Gravity Recovery and Climate Experiment (GRACE) Satellite Mission: A Review. *Remote Sensing*, *10*(6), 829. doi: 10.3390/rs10060829

Freedman, F. R., Pitts, K. L., & Bridger, A. F. C. (2014). Evaluation of CMIP climate model hydrological output for the Mississippi River Basin using GRACE satellite observations. *Journal of Hydrology*, *519*, 3566–3577. doi: 10.1016/j.jhydrol.2014.10.036

Gettelman, A., & Rood, R. B. (2016). *Demystifying Climate Models* (Vol. 2). Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-662-48959-8

GLIMS, & NSIDC. (2005, updated 2020). Global Land Ice Measurements from Space glacier database. Compiled and made available by the international GLIMS community and the National Snow and Ice Data Center, Boulder CO, U.S.A.
doi: 10.7265/N5V98602

Gulev, S. K., Latif, M., Keenlyside, N., Park, W., & Koltermann, K. P. (2013). North Atlantic Ocean control on surface heat flux on multidecadal timescales. *Nature*, *499*(7459), 464–467. doi: 10.1038/nature12268

Güntner, A. (2008). Improvement of Global Hydrological Models Using GRACE Data. *Surveys in Geophysics*, *29*(4), 375–397. doi: 10.1007/s10712-008-9038-y

Han, S.-C., Riva, R., Sauber, J., & Okal, E. (2013). Source parameter inversion for recent great earthquakes from a decade-long observation of global gravity fields. *Journal of Geophysical Research: Solid Earth*, *118*(3), 1240–1267. doi: 10.1002/jgrb.50116

Han, S.-C., Sauber, J., Luthcke, S. B., Ji, C., & Pollitz, F. F. (2008). Implications of postseismic gravity change following the great 2004 Sumatra-Andaman earthquake from the regional harmonic analysis of GRACE intersatellite tracking data. *Journal of Geophysical Research: Solid Earth*, *113*(B11). doi: 10.1029/2008JB005705

Hawkins, E., & Sutton, R. (2009). The Potential to Narrow Uncertainty in Regional Climate Predictions. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1108. doi: 10.1175/2009BAMS2607.1

Heiskanen, W., & Moritz, H. (1967). *Physical Geodesy*. W. H. Freeman.

Horvath, A., Murböck, M., Pail, R., & Horwath, M. (2018). Decorrelation of GRACE Time Variable Gravity Field Solutions Using Full Covariance Information. *Geosciences*, *8*(9), 323. doi: 10.3390/geosciences8090323

Houborg, R., Rodell, M., Li, B., Reichle, R., & Zaitchik, B. F. (2012). Drought indicators based on model-assimilated Gravity Recovery and Climate Experiment (GRACE) terrestrial water storage observations. *Water Resources Research*, *48*(7). doi: https://doi.org/10.1029/2011WR011291

Humphrey, V., & Gudmundsson, L. (2019). GRACE-REC: a reconstruction of climate-driven water storage changes over the last century. *Earth System Science Data*, *11*(3), 1153–1170. doi: https://doi.org/10.5194/essd-11-1153-2019

Humphrey, V., Gudmundsson, L., & Seneviratne, S. I. (2016). Assessing Global Water Storage Variability from GRACE: Trends, Seasonal Cycle, Subseasonal Anomalies and Extremes. *Surveys in Geophysics*, *37*(2), 357–395. doi: 10.1007/s10712-016-9367-1

Humphrey, V., Zscheischler, J., Ciais, P., Gudmundsson, L., Sitch, S., & Seneviratne, S. I. (2018). Sensitivity of atmospheric CO 2 growth rate to observed changes in terrestrial water storage. *Nature*, *560*(7720), 628–631. doi: 10.1038/s41586-018-0424-4

IGSWG. (2016). *NASA/ESA Interagency Gravity Science Working Group: Towards a sustained observing system for mass transport to understand global change and to benefit society.* Doc. Nr.: TUD-IGSWG-2016-01.y.

IPCC. (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. doi: 10.1017/CBO9781107415324

IUGG. (2015). *International Union of Geodesy and Geophysics: Satellite Gravity and Magnetic Mission Constellations.* Resolutions adopted by the council at the XXVI General Assembly Prague, Czech Republic 22. June to 2. July 2015.

Jacob, T., Wahr, J., Pfeffer, W. T., & Swenson, S. (2012). Recent contributions of glaciers and ice caps to sea level rise. *Nature*, *482*(7386), 514–518. doi: 10.1038/nature10847

Jensen, L., Eicker, A., Dobslaw, H., & Pail, R. (2020a). Emerging Changes in Terrestrial Water Storage Variability as a Target for Future Satellite Gravity Missions. *Remote Sensing*, *12*(23), 3898. doi: 10.3390/rs12233898

Jensen, L., Eicker, A., Dobslaw, H., Stacke, T., & Humphrey, V. (2019). Long-Term Wetting and Drying Trends in Land Water Storage Derived From GRACE and CMIP5 Models. *Journal of Geophysical Research: Atmospheres*, *124*(17-18), 9808–9823. doi: 10.1029/2018JD029989

Jensen, L., Eicker, A., Stacke, T., & Dobslaw, H. (2020b). Predictive Skill Assessment for Land Water Storage in CMIP5 Decadal Hindcasts by a Global Reconstruction of GRACE Satellite Data. *Journal of Climate*, *33*(21), 9497–9509. doi: 10.1175/JCLI-D-20-0042.1

Jensen, L., Rietbroek, R., & Kusche, J. (2013). Land water contribution to sea level from GRACE and Jason-1measurements. *Journal of Geophysical Research: Oceans*, *118*(1), 212–226. doi: 10.1002/jgrc.20058

Jha, B., Hu, Z.-Z., & Kumar, A. (2014). SST and ENSO variability and change simulated in historical experiments of CMIP5 models. *Climate Dynamics*, *42*(7), 2113–2124. Retrieved 2021-05-29, from `https://doi.org/10.1007/s00382-013-1803-z` doi: 10.1007/s00382-013-1803-z

Keenlyside, N. S., Latif, M., Jungclaus, J., Kornblueh, L., & Roeckner, E. (2008). Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature*, *453*(7191), 84–88. doi: 10.1038/nature06921

Khaki, M., Hoteit, I., Kuhn, M., Awange, J., Forootan, E., van Dijk, A. I. J. M., ... Pattiaratchi, C. (2017). Assessing sequential data assimilation techniques for integrating GRACE data into a hydrological model. *Advances in Water Resources*, *107*, 301–316. doi: 10.1016/j.advwatres.2017.07.001

Kim, Y.-H., Min, S.-K., Zhang, X., Sillmann, J., & Sandstad, M. (2020). Evaluation of the CMIP6 multi-model ensemble for climate extreme indices. *Weather and Climate Extremes*, *29*, 100269. doi: 10.1016/j.wace.2020.100269

Kishore, P., Basha, G., Venkat Ratnam, M., Velicogna, I., Ouarda, T. B. M. J., & Narayana Rao, D. (2016). Evaluating CMIP5 models using GPS radio occultation COSMIC temperature in UTLS region during 2006–2013: twenty-first century projection and trends. *Climate Dynamics*, *47*(9), 3253–3270. doi: 10.1007/s00382-016-3024-8

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., & Meehl, G. A. (2010). Challenges in Combining Projections from Multiple Climate Models. *Journal of Climate*, *23*(10), 2739–2758. doi: 10.1175/2009JCLI3361.1

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., & Eyring, V. (2017). A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters*, *44*(4), 1909–1918. doi: https://doi.org/10.1002/2016GL072012

Koelling, J., Send, U., & Lankhorst, M. (2020). Decadal Strengthening of Interior Flow of North Atlantic Deep Water Observed by GRACE Satellites. *Journal of Geophysical Research: Oceans*, *125*(11), e2020JC016217. doi: https://doi.org/10.1029/2020JC016217

Kornfeld, R. P., Arnold, B. W., Gross, M. A., Dahya, N. T., Klipstein, W. M., Gath, P. F., & Bettadpur, S. (2019). GRACE-FO: The Gravity Recovery and Climate Experiment Follow-On Mission. *Journal of Spacecraft and Rockets*, *56*(3), 931–951. doi: 10.2514/1.A34326

Kusche, J. (2007). Approximate decorrelation and non-isotropic smoothing of time-variable GRACE-type gravity field models. *Journal of Geodesy*, *81*(11), 733–749. doi: 10.1007/s00190-007-0143-3

Kusche, J., Eicker, A., Forootan, E., Springer, A., & Longuevergne, L. (2016). Mapping probabilities of extreme continental water storage changes from space gravimetry. *Geophysical Research Letters*, *43*(15), 8026–8034. doi: 10.1002/2016GL069538

Kvas, A., Behzadpour, S., Ellmer, M., Klinger, B., Strasser, S., Zehentner, N., & Mayer-Gürr, T. (2019). ITSG-Grace2018: Overview and Evaluation of a New GRACE-Only Gravity Field Time Series. *Journal of Geophysical Research: Solid Earth*, *124*(8), 9332–9344. doi: 10.1029/2019JB017415

Laepple, T., & Huybers, P. (2014). Ocean surface temperature variability: Large model–data differences at decadal and longer periods. *Proceedings of the National Academy of Sciences*, *111*(47), 16682–16687. doi: 10.1073/pnas.1412077111

Lambeck, K. (1988). *Geophysical geodesy : the slow deformations of the earth*. Oxford [Oxfordshire] : Clarendon Press ; New York : Oxford University Press.

Landerer, F. W., Flechtner, F. M., Save, H., Webb, F. H., Bandikova, T., Bertiger, W. I., ... Yuan, D.-N. (2020). Extending the Global Mass Change Data Record: GRACE Follow-On Instrument and Science Data Performance. *Geophysical Research Letters*, *47*(12), e2020GL088306. doi: 10.1029/2020GL088306

Landerer, F. W., Gleckler, P. J., & Lee, T. (2014). Evaluation of CMIP5 dynamic sea surface height multi-model simulations against satellite observations. *Climate Dynamics*, *43*(5), 1271–1283. doi: 10.1007/s00382-013-1939-x

Landerer, F. W., Wiese, D. N., Bentel, K., Boening, C., & Watkins, M. M. (2015). North Atlantic meridional overturning circulation variations from GRACE ocean bottom pressure anomalies. *Geophysical Research Letters*, *42*(19), 8114–8121. doi: https://doi.org/10.1002/2015GL065730

Lauer, A., Eyring, V., Righi, M., Buchwitz, M., Defourny, P., Evaldsson, M., ... Willén, U. (2017). Benchmarking CMIP5 models with a subset of ESA CCI Phase 2 data using the ESMValTool. *Remote Sensing of Environment*, *203*, 9–39. doi: 10.1016/j.rse.2017.01.007

Lenczuk, A., Leszczuk, G., Klos, A., Kosek, W., & Bogusz, J. (2020). Study on the inter-annual hydrology-induced deformations in Europe using GRACE and hydrological models. *Journal of Applied Geodesy*, *14*(4), 393–403. doi: 10.1515/jag-2020-0017

Lettenmaier, D. P., & Famiglietti, J. S. (2006). Hydrology: Water from on high. *Nature*, *444*(7119), 562–563. doi: 10.1038/444562a

Liang, Y.-C., Lo, M.-H., Lan, C.-W., Seo, H., Ummenhofer, C. C., Yeager, S., ... Steffen, J. D. (2020). Amplified seasonal cycle in hydroclimate over the Amazon river basin and its plume region. *Nature Communications*, *11*(1), 4390. doi: 10.1038/s41467-020-18187-0

Liepert, B. G., & Lo, F. (2013). CMIP5 update of 'Inter-model variability and biases of the global water cycle in CMIP3 coupled climate models'. *Environmental Research Letters*, *8*(2), 029401. doi: 10.1088/1748-9326/8/2/029401

Liu, C., Allan, R. P., & Huffman, G. J. (2012). Co-variation of temperature and precipitation in CMIP5 models and satellite observations. *Geophysical Research Letters*, *39*(13). doi: https://doi.org/10.1029/2012GL052093

Loomis, B. D., Rachlin, K. E., & Luthcke, S. B. (2019). Improved Earth Oblateness Rate Reveals Increased Ice Sheet Losses and Mass-Driven Sea Level Rise. *Geophysical Research Letters*, *46*(12), 6910–6917. doi: https://doi.org/10.1029/2019GL082929

Lorenz, E. N. (1963). Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*, *20*(2), 130–141. doi: 10.1175/1520-0469(1963)020⟨0130:DNF⟩2.0.CO;2

Luković, J., Chiang, J. C. H., Blagojević, D., & Sekulić, A. (2021). A Later Onset of the Rainy Season in California. *Geophysical Research Letters*, *48*(4), e2020GL090350. doi: https://doi.org/10.1029/2020GL090350

Mayer-Gürr, T., Behzadpur, S., Ellmer, M., Kvas, A., Klinger, B., Strasser, S., & Zehentner, N. (2018). *ITSG-Grace2018 - Monthly, Daily and Static Gravity Field Solutions from GRACE.* GFZ Data Services. doi: 10.5880/ICGEM.2018.003

Meehl, G. A., Goddard, L., Boer, G., Burgman, R., Branstator, G., Cassou, C., ... Yeager, S. (2013). Decadal Climate Prediction: An Update from the Trenches. *Bulletin of the American Meteorological Society*, *95*(2), 243–267. doi: 10.1175/BAMS-D-12-00241.1

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., ... Stockdale, T. (2009). Decadal Prediction. *Bulletin of the American Meteorological Society*, *90*(10), 1467–1485. doi: 10.1175/2009BAMS2778.1

Mehran, A., AghaKouchak, A., & Phillips, T. J. (2014). Evaluation of CMIP5 continental precipitation simulations relative to satellite-based gauge-adjusted observations. *Journal of Geophysical Research: Atmospheres*, *119*(4), 1695–1707. doi: https://doi.org/10.1002/2013JD021152

Mehrotra, R., Sharma, A., Bari, M., Tuteja, N., & Amirthanathan, G. (2014). An assessment of CMIP5 multi-model decadal hindcasts over Australia from a hydrological viewpoint. *Journal of Hydrology*, *519*(Part D), 2932–2951. doi: 10.1016/j.jhydrol.2014.07.053

Moreira, L., Cazenave, A., & Palanisamy, H. (2021). Influence of interannual variability in estimating the rate and acceleration of present-day global mean sea level. *Global and Planetary Change*, *199*, 103450. doi: 10.1016/j.gloplacha.2021.103450

Mu, D., Yan, H., Feng, W., & Peng, P. (2017). GRACE leakage error correction with regularization technique: case studies in Greenland and Antarctica. *Geophysical Journal International*, *208*(3), 1775–1786. doi: 10.1093/gji/ggw494

Murböck, M., Pail, R., Daras, I., & Gruber, T. (2014). Optimal orbits for temporal gravity recovery regarding temporal aliasing. *Journal of Geodesy*, *88*(2), 113–126. doi: 10.1007/s00190-013-0671-y

Ni, S., Chen, J., Wilson, C. R., Li, J., Hu, X., & Fu, R. (2018). Global Terrestrial Water Storage Changes and Connections to ENSO Events. *Surveys in Geophysics*, *39*(1), 1–22. doi: 10.1007/s10712-017-9421-7

Pail, R., Bingham, R., Braitenberg, C., Dobslaw, H., Eicker, A., Güntner, A., . . . IUGG Expert Panel (2015). Science and User Needs for Observing Global Mass Transport to Understand Global Change and to Benefit Society. *Surveys in Geophysics*, *36*(6), 743–772. doi: 10.1007/s10712-015-9348-9

Palmer, T., Buizza, R., Hagedorn, R., Lawrence, A., Leutbecher, M., & Smith, L. (2006). Ensemble prediction: A pedagogical perspective. *ECMWF*. doi: 10.21957/AB129056EW

Pennell, C., & Reichler, T. (2011). On the Effective Number of Climate Models. *Journal of Climate*, *24*(9), 2358–2367. doi: 10.1175/2010JCLI3814.1

Phillips, T., Nerem, R. S., Fox-Kemper, B., Famiglietti, J. S., & Rajagopalan, B. (2012). The influence of ENSO on global terrestrial water storage using GRACE. *Geophysical Research Letters*, *39*(16). doi: 10.1029/2012GL052495

Purkhauser, A. F., & Pail, R. (2019). Next generation gravity missions: near-real time gravity field retrieval strategy. *Geophysical Journal International*, *217*(2), 1314–1333. doi: 10.1093/gji/ggz084

Randall, D. A. (2017). *An Introduction to Numerical Modeling of the Atmosphere*. Retrieved 2021-05-31, from `http://kiwi.atmos.colostate.edu/rr/group.html`

Reager, J. T., Gardner, A. S., Famiglietti, J. S., Wiese, D. N., Eicker, A., & Lo, M.-H. (2016). A decade of sea level rise slowed by climate-driven hydrology. *Science*, *351*(6274), 699–703. doi: 10.1126/science.aad8386

Reager, J. T., Thomas, B. F., & Famiglietti, J. S. (2014). River basin flood potential inferred using GRACE gravity observations at several months lead time. *Nature Geoscience*, *7*(8), 588–592. doi: 10.1038/ngeo2203

Riahi, K., van Vuuren, D. P., Kriegler, E., Edmonds, J., O'Neill, B. C., Fujimori, S., . . . Tavoni, M. (2017). The Shared Socioeconomic Pathways and their energy, land use, and greenhouse gas emissions implications: An overview. *Global Environmental Change*, *42*, 153–168. doi: 10.1016/j.gloenvcha.2016.05.009

Rietbroek, R., Brunnabend, S.-E., Kusche, J., Schröter, J., & Dahle, C. (2016). Revisiting the contemporary sea-level budget on global and regional scales. *Proceedings of the National Academy of Sciences*, *113*(6), 1504–1509. doi: 10.1073/pnas.1519132113

Rodell, M., Famiglietti, J. S., Wiese, D. N., Reager, J. T., Beaudoing, H. K., Landerer, F. W., & Lo, M.-H. (2018). Emerging trends in global freshwater availability. *Nature*, *557*(7707), 651. doi: 10.1038/s41586-018-0123-1

Rodell, M., Velicogna, I., & Famiglietti, J. S. (2009). Satellite-based estimates of groundwater depletion in India. *Nature*, *460*(7258), 999–1002. doi: 10.1038/nature08238

Roman, J. A., Knuteson, R. O., Ackerman, S. A., Tobin, D. C., & Revercomb, H. E. (2012). Assessment of Regional Global Climate Model Water Vapor Bias and Trends Using Precipitable Water Vapor (PWV) Observations from a Network of Global Positioning Satellite (GPS) Receivers in the U.S. Great Plains and Midwest. *Journal of Climate*, *25*(16), 5471–5493. doi: 10.1175/JCLI-D-11-00570.1

Salerno, J., Diem, J. E., Konecky, B. L., & Hartter, J. (2019). Recent intensification of the seasonal rainfall cycle in equatorial Africa revealed by farmer perceptions, satellite-based estimates, and ground-based station measurements. *Climatic Change*, *153*(1), 123–139. doi: 10.1007/s10584-019-02370-4

Sanderson, B. M., Knutti, R., & Caldwell, P. (2015). A Representative Democracy to Reduce Interdependency in a Multimodel Ensemble. *Journal of Climate*, *28*(13), 5171–5194. doi: 10.1175/JCLI-D-14-00362.1

Sasgen, I., van den Broeke, M., Bamber, J. L., Rignot, E., Sørensen, L. S., Wouters, B., ... Simonsen, S. B. (2012). Timing and origin of recent regional ice-mass loss in Greenland. *Earth and Planetary Science Letters*, *333-334*, 293–303. doi: 10.1016/j.epsl.2012.03.033

Save, H., Bettadpur, S., & Tapley, B. D. (2016). High-resolution CSR GRACE RL05 mascons. *Journal of Geophysical Research: Solid Earth*, *121*(10), 7547–7569. doi: https://doi.org/10.1002/2016JB013007

Scanlon, B. R., Zhang, Z., Rateb, A., Sun, A., Wiese, D., Save, H., ... Reedy, R. C. (2019). Tracking Seasonal Fluctuations in Land Water Storage Using Global Models and GRACE Satellites. *Geophysical Research Letters*, *46*(10), 5254–5264. doi: 10.1029/2018GL081836

Scanlon, B. R., Zhang, Z., Save, H., Sun, A. Y., Müller Schmied, H., van Beek, L. P. H., ... Bierkens, M. F. P. (2018). Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. *Proceedings of the National Academy of Sciences*, *115*(6), E1080–E1089. doi: 10.1073/pnas.1704665115

Schwalm, C. R., Glendon, S., & Duffy, P. B. (2020). RCP8.5 tracks cumulative CO2 emissions. *Proceedings of the National Academy of Sciences*, *117*(33), 19656–19657. doi: 10.1073/pnas.2007117117

Shu, Q., Wang, Q., Song, Z., Qiao, F., Zhao, J., Chu, M., & Li, X. (2020). Assessment of Sea Ice Extent in CMIP6 With Comparison to Observations and CMIP5. *Geophysical Research Letters*, 47(9), e2020GL087965. doi: https://doi.org/10.1029/2020GL087965

Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W., & Bronaugh, D. (2013). Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. *Journal of Geophysical Research: Atmospheres*, *118*(4), 1716–1733. doi: 10.1002/jgrd.50203

Smith, D. M., Eade, R., Scaife, A. A., Caron, L.-P., Danabasoglu, G., DelSole, T. M., ... Yang, X. (2019). Robust skill of decadal climate predictions. *npj Climate and Atmospheric Science*, *2*(1), 1–10. doi: 10.1038/s41612-019-0071-y

Steffen, H., Wu, P., & Wang, H. (2010). Determination of the Earth's structure in Fennoscandia from GRACE and implications for the optimal post-processing of GRACE data. *Geophysical Journal International*, *182*(3), 1295–1310. doi: 10.1111/j.1365-246X.2010.04718.x

Su, F., Duan, X., Chen, D., Hao, Z., & Cuo, L. (2013). Evaluation of the Global Climate Models in the CMIP5 over the Tibetan Plateau. *Journal of Climate*, *26*(10), 3187–3208. doi: 10.1175/JCLI-D-12-00321.1

Sun, Y., Riva, R., & Ditmar, P. (2016). Optimizing estimates of annual variations and trends in geocenter motion and J_2 from a combination of GRACE data and geophysical models. *Journal of Geophysical Research: Solid Earth*, *121*(11), 8352–8370. doi: 10.1002/2016JB013073

Swenson, S., Chambers, D., & Wahr, J. (2008). Estimating geocenter variations from a combination of GRACE and ocean model output. *Journal of Geophysical Research: Solid Earth*, *113*(B8). doi: https://doi.org/10.1029/2007JB005338

Swenson, S., & Wahr, J. (2006). Post-processing removal of correlated errors in GRACE data. *Geophysical Research Letters*, *33*(8). doi: https://doi.org/10.1029/2005GL025285

Tapley, B. D., Bettadpur, S., Watkins, M., & Reigber, C. (2004). The gravity recovery and climate experiment: Mission overview and early results. *Geophysical Research Letters*, *31*(9), L09607. doi: 10.1029/2004GL019920

Tapley, B. D., Watkins, M. M., Flechtner, F., Reigber, C., Bettadpur, S., Rodell, M., . . . Velicogna, I. (2019). Contributions of GRACE to understanding climate change. *Nature Climate Change*. doi: 10.1038/s41558-019-0456-2

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2011). An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. doi: 10.1175/BAMS-D-11-00094.1

Tebaldi, C., & Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *365*(1857), 2053–2075. doi: 10.1098/rsta.2007.2076

Tian, B., Fetzer, E. J., Kahn, B. H., Teixeira, J., Manning, E., & Hearty, T. (2013). Evaluating CMIP5 models using AIRS tropospheric air temperature and specific humidity climatology. *Journal of Geophysical Research: Atmospheres*, *118*(1), 114–134. doi: https://doi.org/10.1029/2012JD018607

Tiwari, V. M., Wahr, J., & Swenson, S. (2009). Dwindling groundwater resources in northern India, from satellite gravity observations. *Geophysical Research Letters*, *36*(18), L18401. doi: 10.1029/2009GL039401

Turner, J., Bracegirdle, T. J., Phillips, T., Marshall, G. J., & Hosking, J. S. (2013). An Initial Assessment of Antarctic Sea Ice Extent in the CMIP5 Models. *Journal of Climate*, *26*(5), 1473–1484. doi: 10.1175/JCLI-D-12-00068.1

Velicogna, I., & Wahr, J. (2013). Time-variable gravity observations of ice sheet mass balance: Precision and limitations of the GRACE satellite data. *Geophysical Research Letters*, *40*(12), 3055–3063. doi: 10.1002/grl.50527

Vignesh, P. P., Jiang, J. H., Kishore, P., Su, H., Smay, T., Brighton, N., & Velicogna, I. (2020). Assessment of CMIP6 Cloud Fraction and Comparison with Satellite Observations. *Earth and Space Science*, *7*(2), e2019EA000975. doi: https://doi.org/10.1029/2019EA000975

Wanders, N., & Wada, Y. (2015). Decadal predictability of river discharge with climate oscillations over the 20th and early 21st century. *Geophysical Research Letters*, *42*(24), 10,689–10,695. doi: https://doi.org/10.1002/2015GL066929

Watkins, M. M., Wiese, D. N., Yuan, D.-N., Boening, C., & Landerer, F. W. (2015). Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons. *Journal of Geophysical Research: Solid Earth*, *120*(4), 2648–2671. doi: https://doi.org/10.1002/2014JB011547

Wiese, D. N., Visser, P., & Nerem, R. S. (2011). Estimating low resolution gravity fields at short time intervals to reduce temporal aliasing errors. *Advances in Space Research*, *48*(6), 1094–1107. doi: 10.1016/j.asr.2011.05.027

Yuan, X., & Zhu, E. (2018). A First Look at Decadal Hydrological Predictability by Land Surface Ensemble Simulations. *Geophysical Research Letters*, *45*(5), 2362–2369. doi: 10.1002/2018GL077211

Zhang, L., Dobslaw, H., Dahle, C., Sasgen, I., & Thomas, M. (2016). Validation of MPI-ESM Decadal Hindcast Experiments with Terrestrial Water Storage Variations as Observed by the GRACE Satellite Mission. *Meteorologische Zeitschrift*, 685–694. doi: 10.1127/metz/2015/0596

Zhang, L., Dobslaw, H., Stacke, T., Güntner, A., Dill, R., & Thomas, M. (2017). Validation of terrestrial water storage variations as simulated by different global numerical models with GRACE satellite observations. *Hydrol. Earth Syst. Sci.*, *21*(2), 821–837. doi: 10.5194/hess-21-821-2017

Śliwińska, J., Wińska, M., & Nastula, J. (2019). Terrestrial water storage variations and their effect on polar motion. *Acta Geophysica*, *67*(1), 17–39. doi: 10.1007/s11600-018-0227-x

Acronyms

AR Assessment Report.

CMIP Coupled Model Intercomparison Project.
COST-G Combination Service for Time-variable Gravity Fields.
CSR Center for Space Research.

DLR German Aerospace Center.

ECDF empirical cumulative density function.
ECV Essential Climate Variable.
ENSO El Niño Southern Oscillation.
ESA European Space Agency.
ESM Earth System Model.
EWH equivalent water height.

GCM General Circulation Model.
GCOS Global Climate Observing System.
GFZ German Research Centre for Geosciences.
GHG greenhouse gas.
GIA glacial isostatic adjustment.
GLIMS Global Land Ice Measurements from Space.
GNSS Global Navigation Satellite System.
GRACE Gravity Recovery and Climate Experiment.
GRACE-FO GRACE Follow-On.

IGFS International Gravity Field Service.
IPCC Intergovernmental Panel on Climate Change.
ITSG Institute of Geodesy, Working Group Theoretical Geodesy and Satellite Geodesy.

JPL Jet Propulsion Laboratory.

LRI laser-ranging interferometer.

MMMed multi-model median.
mrso total soil moisture content.
mTWS modeled terrestrial water storage.

NASA National Aeronautics and Space Administration.
NGGM Next Generation Gravity Mission.

RCP Representative Concentration Pathway.
RMS root mean square.

RMSD root mean squared deviation.

SMM soil moisture memory.
snw surface snow amount.
SSP Shared Socioeconomic Pathway.

TWS terrestrial water storage.

WCRP World Climate Research Programme.
WMO World Meteorological Organization.

List of Figures

List of Tables

Appendix

Appendix A.

Publications

A.1. Long-Term Wetting and Drying Trends in Land Water Storage Derived From GRACE and CMIP5 Models

Reference

Jensen, L., Eicker, A., Dobslaw, H., Stacke, T., Humphrey, V. (2019). Long-Term Wetting and Drying Trends in Land Water Storage Derived From GRACE and CMIP5 Models. Journal of Geophysical Research: Atmospheres 124, 9808–9823.
https://doi.org/10.1029/2018JD029989

Abstract

Coupled climate models participating in the CMIP5 (Coupled Model Intercomparison Project Phase 5) exhibit a large intermodel spread in the representation of long-term trends in soil moisture and snow in response to anthropogenic climate change. We evaluate long-term (January 1861 to December 2099) water storage trends from 21 CMIP5 models against observed trends in terrestrial water storage (TWS) obtained from 14 years (April 2002 to August 2016) of the GRACE (Gravity Recovery And Climate Experiment) satellite mission. This is complicated due to the incomplete representation of TWS in CMIP5 models and interannual climate variability masking long-term trends in observations. We thus evaluate first the spread in projected trends among CMIP5 models and identify regions of broad model consensus. Second, we assess the extent to which these projected trends are already present during the historical period (January 1861 to August 2016) and thus potentially detectable in observational records available today. Third, we quantify the degree to which 14-year tendencies can be expected to represent long-term trends, finding that regional long-term trends start to emerge from interannual variations after just 14 years while stable global trend patterns are detectable after 30 years. We classify regions of strong model consensus into areas where (1) climate-related TWS changes are supported by the direction of GRACE trends, (2) mismatch of trends hints at possible model deficits, (3) the short observation time span and/or anthropogenic influences prevent reliable conclusions about long-term wetting or drying. We thereby demonstrate the value of satellite observations of water storage to further constrain the response of the terrestrial water cycle to climate change.

Declaration of own contribution

Table A.1.: Contribution to Paper No. 1.

Involved in	Estimated contribution
Ideas and conceptual design	85%
Computation and results	100%
Analysis and interpretation	85%
Manuscript, figures and tables	90%
Total	90%

Confirmation of Co-Authors

I hereby confirm the correctness of the declaration of the contribution of Laura Jensen for Paper No. 1 in Table A.1:

26.05.21

Annette Eicker
(HCU Hamburg)
Date

Henryk
Dobslaw
Digital unterschrieben von
Henryk Dobslaw
Datum: 2021.05.26 20:26:59
+02'00'

Henryk Dobslaw
(GFZ Potsdam)
Date

Tobias Stacke
28.05.21

Tobias Stacke
(HZG Geesthacht)
Date

28.05.21

Vincent Humphrey
(JPL California)
Date

AGU100 ADVANCING EARTH AND SPACE SCIENCE

JGR Atmospheres

RESEARCH ARTICLE
10.1029/2018JD029989

Long-Term Wetting and Drying Trends in Land Water Storage Derived From GRACE and CMIP5 Models

L. Jensen[1], A. Eicker[1], H. Dobslaw[2], T. Stacke[3], and V. Humphrey[4]

[1]Geodesy and Geoinformatics, HafenCity University Hamburg, Hamburg, Germany, [2]Helmholtz Centre Potsdam, German Research Centre for Geosciences (GFZ), Potsdam, Germany, [3]Max Planck Institute for Meteorology, Hamburg, Germany, [4]Institute for Atmospheric and Climate Science, ETH Zürich, Zürich, Switzerland

Key Points:
- By comparing terrestrial water storage trends from CMIP5 models and GRACE satellite data, we identify hot spot regions of wetting and drying
- Model analysis reveals that regional long-term trends start to emerge from interannual variations after 14 years
- Large model spread in water storage trends demonstrates importance of GRACE to constrain response of the water cycle to climate change

Correspondence to:
L. Jensen,
laura.jensen@hcu-hamburg.de

Citation:
Jensen, L., Eicker, A., Dobslaw, H., Stacke, T., & Humphrey, V. (2019). Long-term wetting and drying trends in land water storage derived from GRACE and CMIP5 models. *Journal of Geophysical Research: Atmospheres*, *124*, 9808–9823. https://doi.org/10.1029/2018JD029989

Received 14 NOV 2018
Accepted 17 AUG 2019
Accepted article online 29 AUG 2019
Published online 3 SEP 2019

Abstract

Coupled climate models participating in the CMIP5 (Coupled Model Intercomparison Project Phase 5) exhibit a large intermodel spread in the representation of long-term trends in soil moisture and snow in response to anthropogenic climate change. We evaluate long-term (January 1861 to December 2099) water storage trends from 21 CMIP5 models against observed trends in terrestrial water storage (TWS) obtained from 14 years (April 2002 to August 2016) of the GRACE (Gravity Recovery And Climate Experiment) satellite mission. This is complicated due to the incomplete representation of TWS in CMIP5 models and interannual climate variability masking long-term trends in observations. We thus evaluate first the spread in projected trends among CMIP5 models and identify regions of broad model consensus. Second, we assess the extent to which these projected trends are already present during the historical period (January 1861 to August 2016) and thus potentially detectable in observational records available today. Third, we quantify the degree to which 14-year tendencies can be expected to represent long-term trends, finding that regional long-term trends start to emerge from interannual variations after just 14 years while stable global trend patterns are detectable after 30 years. We classify regions of strong model consensus into areas where (1) climate-related TWS changes are supported by the direction of GRACE trends, (2) mismatch of trends hints at possible model deficits, (3) the short observation time span and/or anthropogenic influences prevent reliable conclusions about long-term wetting or drying. We thereby demonstrate the value of satellite observations of water storage to further constrain the response of the terrestrial water cycle to climate change.

1. Introduction

The terrestrial branch of the global water cycle is an important component of the Earth's coupled climate system: Water available in the soil critically determines biomass production that effectively takes up carbon dioxide from the atmosphere and thus constitutes the land cover and consequently also the albedo of the Earth's surface. The availability of water at the surface influences the rate of evapotranspiration and thereby the amount of latent heat absorbed by the atmosphere locally and advected to distant regions along with the tropospheric winds; and water in the form of snow cover thermally isolates the soil from the air above it. The accurate representation of the terrestrial water dynamics and its various feedbacks to the atmospheric water, energy and carbon cycles is thus critically important for interactively coupled global numerical climate models that are used to infer information about the current state and the future evolution of the Earth's climate conditions (Trenberth, 2010).

Due to their direct effect on the availability of freshwater resources, investigating climate change impacts on the global water cycle is of great societal relevance. Changes in terrestrial water storage (TWS) might reflect long-term wetting or drying in various regions of the world, and the identification of such regions is of substantial importance for water resources management. However, coupled climate models used to predict future climatic conditions still exhibit a spread in the representation of long-term trends in soil moisture and other land water related variables (Guo & Dirmeyer, 2006; Figure 12.23 in Berg et al., 2017; Collins et al., 2013; Yuan & Quiring, 2017).

Comparing the output of numerical models with observations is crucial to demonstrate their reliability and to test predictive capacities, but measurements of water storage changes are difficult to obtain. A classical approach for the determination of TWS at basin scale is the integration of the water balance equation (precipitation minus evapotranspiration minus runoff), see Rodell et al. (2004). However, this is challenging on a global scale, since streamflow measurements are sparse and evapotranspiration is generally difficult to measure (Wartenburger et al., 2018). Especially, trends in water storage cannot be recovered well by this method due to biases in the water flux observations (Hirschi & Seneviratne, 2017).

Complementary to conventional meteorologic observations of atmospheric water fluxes, the satellite mission Gravity Recovery And Climate Experiment (GRACE; Tapley et al., 2004) in operation from 2002 to 2017 allowed for the first time the observation of water storage changes with global coverage from space. By evaluating relative distance changes between two spacecraft at very low altitudes of 400–500 km, time variations in the Earth's gravity field are mapped that can be unambiguously related to changes in TWS. Due to the indirect observation concept, GRACE essentially senses water mass anomalies independently of their surface exposure and thus integrates all mass changes vertically from the surface down to the deepest aquifers. This unique capability of the gravimetric method makes GRACE highly complementary to alternative radiometric satellite techniques of soil moisture remote sensing that are only sensitive to changes in the top few centimeters of soil (Dorigo et al., 2015). GRACE mission data have been used in various hydrometeorological applications, for example, Famiglietti and Rodell (2013), and it is rated among the top five priorities of the future Earth observation capacity by the most recent National Aeronautics and Space Administration decadal survey (Committee on the Decadal Survey for Earth Science and Applications from Space et al., 2018). The successor mission GRACE-FO (Follow On), launched in May 2018, is expected to continue this important observational record over the next decades (Flechtner et al., 2016), which will facilitate the separation between interannual variability and long-term climatological trends in TWS. Because the limited time span of GRACE data makes the identification of climate-related signals still challenging, this study aims to investigate how GRACE TWS trends could (and should) be compared to model-derived trends.

TWS as observed with GRACE has already been used to validate both global hydrological models (Döll et al., 2014; Eicker et al., 2014; Güntner, 2008; Syed et al., 2008) and land surface models (Scanlon et al., 2018; Zhang et al., 2017) which are driven by a prescribed meteorological forcing. In this study, we focus on interactively coupled Earth System Models (ESMs) participating in CMIP5 (Coupled Model Intercomparison Project Phase 5, Taylor et al., 2011). Comparing GRACE trends with long-term coupled climate model projections is challenging in mainly two aspects: (i) In contrast to GRACE TWS (i.e., the full integrated water column, including all water reservoirs), TWS in the models is reflected typically only by means of snow storage and soil moisture. The representation of the latter critically depends on the depth of the soil column and the number of vertical layers considered. In particular, current ESMs do not explicitly simulate groundwater storage changes. As groundwater-surface interactions play an important role in the global hydrological cycle, this poses an additional source of uncertainty in long-term model projections of wetting and drying. (ii) Coupled runs in CMIP5 starting from preindustrial conditions and extending over the whole historical period until the present day are forced with temporally variable solar radiation, aerosols, CO_2 concentrations, and land use. Those experiments are thus expected to reproduce the climate variability in a statistical sense only. As a result, different realizations of the interannual and decadal climate variability are superimposed over the climatological trends so that a direct comparison with the 14-year GRACE TWS time series only has limited explanatory power. While a regional study for the Mississippi Basin (Freedman et al., 2014) showed reasonably good agreement for the annual amplitude of GRACE data and a subset of CMIP5 models, Fasullo et al. (2016) found the trends from historical CESM1-CAM5 runs compared to GRACE to be dominated by internal variability rather than by the forced response. Different drivers of TWS trends observed by GRACE were investigated by Rodell et al. (2018), who also made use of CMIP5 model precipitation projections to attribute wetting and drying tendencies in some regions to climate-driven precipitation changes. To our knowledge, an extensive global comparison of soil moisture and snow trends modeled over more than two centuries (in the following referred to as bicentennial) against GRACE observations has never been conducted with an ensemble of models such as CMIP5.

In response to these challenges, we focus in this study in particular on the correspondence of bicentennial trends in TWS as simulated by the majority of CMIP5 models and TWS tendencies as observed by GRACE and investigate regions of agreement and disagreement on wetting or drying trends in models and satellite observations.

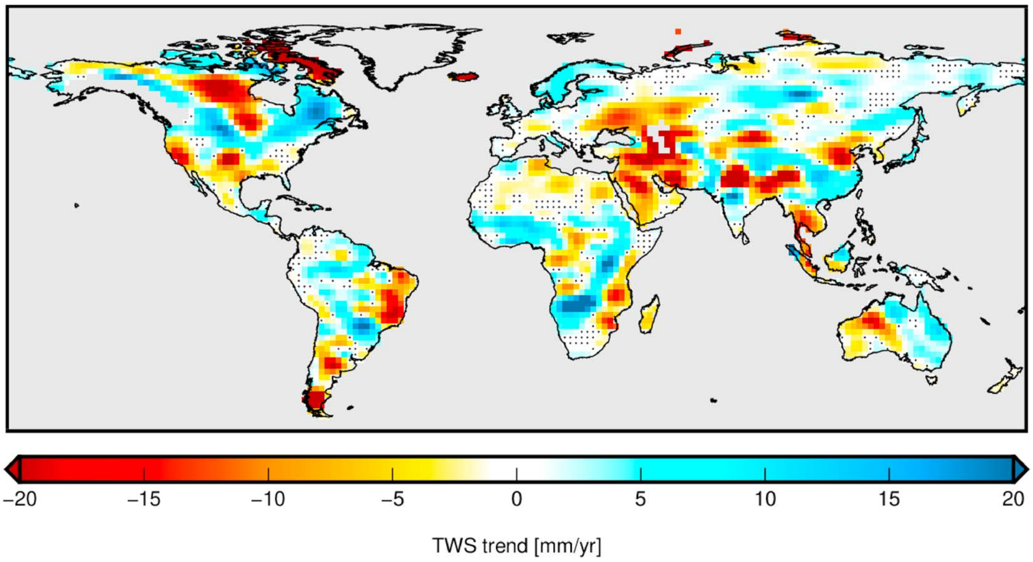


Figure 1. TWS trends from ITSG-Grace2018s (preliminary) for the time span April 2002 to August 2016 (without Greenland, Svalbard, Gulf Coast of Alaska, and Antarctica). Stippling indicates regions with nonsignificant trends ($\alpha = 0.05$).

This paper is structured as follows: First, we compute global maps of TWS trends from GRACE data (section 2) and CMIP5 models (section 3) together with an evaluation of the variability among different models and within historical and future time spans. We estimate the influence of the different time series lengths for GRACE and models by means of two model studies using model TWS tendencies from time periods ranging from 14 to more than 200 years (section 4). The TWS trend maps from GRACE and CMIP5 models are subsequently compared (section 5). Next, we investigate hot spot and noncompliance regions of wetting and drying trends regarding their uncertainty (section 6), which might be caused by model deficits or natural interannual variability and human impacts affecting the GRACE-derived trends. Section 7 summarizes the results and addresses future work.

2. TWS Trends From GRACE Data

To obtain a global grid of observed TWS trends we use the ITSG-Grace2018s trend Level 2 data (Mayer-Gürr et al., 2018), which was obtained from estimating a long-term mean gravity field model together with linear trend and annual cycle from all available GRACE Level 1B RL03 data. The ITSG-Grace2018s trend used here is a preliminary version containing Level 1B data of the time span April 2002 to August 2016 (~14 years). It will be updated once the complete time series of Level 1B RL03 data (April 2002 to June 2017) is available. However, for the trend only minor changes are expected by extending the time series by less than 1 year.

The spherical harmonic coefficients (Level 2) of the trend in gravitational potential are given up to degree $n_{\max} = 120$ and are postprocessed as follows: The effect of geocenter motion is taken into account by augmenting the GRACE data with the linear trends of degree 1 harmonic coefficients provided by Swenson et al. (2008). The zonal Δc_{20} trend coefficient is replaced using a result from Satellite Laser Ranging (Cheng et al., 2013). To reduce mass trends originating from glacial isostatic adjustment (GIA), we subtract a model from A et al. (2013) and to mitigate the effect of correlated noise a DDK4 filter (Kusche, 2007) is applied. We calculate the TWS trend on a 2° × 2° geographical grid (Figure 1) according to

$$tws(\lambda, \theta) = \frac{M}{4\pi R^2 \rho_w} \sum_{n=1}^{n_{\max}} \sum_{m=-n}^{n} \frac{(2n+1)}{(1+k'_n)} \Delta c_{nm} Y_{nm}(\lambda, \theta) \tag{1}$$

where λ and θ denote the spherical coordinates, M and R are the mass and the radius of the Earth, $\rho_w = 1,000$ kg/m^3 is the density of water, k'_n denote the Load Love Numbers (Lambeck, 1988), Δc_{nm} are the filtered spherical harmonic coefficients of the gravitational potential, and $Y_{nm}(\lambda, \theta)$ are the surface spherical harmonic functions. Corresponding standard deviations of the TWS trends are obtained by variance propagation from realistic error assumptions provided with the Δc_{nm} coefficients of the ITSG-Grace2018s trend.

The significance of the trend can be tested with a parameter test. The estimated trend divided by its estimated standard deviation is compared to the critical value of the normal distribution for a certain significance level $1 - \alpha$ which we set to 95% in this study. Generally, the reliability of trends from GRACE is high. Among the solutions of different GRACE processing centers trends over the same time period are very similar (Scanlon et al., 2018), even if a different representation (mascons instead of spherical harmonics) is chosen. Thus, selecting another GRACE solution (e.g., from JPL or CSR) does not alter the findings of our study (not shown).

GRACE-derived trends might not originate purely from TWS changes everywhere, as residual tectonic effects from GIA (Caron et al., 2018), postseismic deformation after large earthquakes (Han et al., 2008, 2010), or residual atmospheric mass variability (Fagiolini et al., 2015) can overlay TWS trends. Furthermore, leakage of signal into neighboring grid cells due to filtering and residual noise that could not be removed during filtering might also distort TWS trends.

As the GRACE TWS trends are only calculated from 14 years of data, the results can be dominated by low-frequency climate variability related to El Niño–Southern Oscillation (Ni et al., 2018; Phillips et al., 2012), the solar cycle (Bhattacharyya & Narasimha, 2005), the quasi-biennial oscillation and other coupled climate modes (Gray et al., 2018), and episodic events as volcanic eruptions (Iles et al., 2013), which may either conceal the long-term trend or produce a spurious transient trend. Approaches to reduce these interannual variabilities in the GRACE record are currently being discussed (e.g., Eicker et al., 2016).

3. TWS Trends From CMIP5 Model Data

As CMIP5 models do not provide a standard output variable for total water storage, we use the sum of total soil moisture content (mrso) and surface snow amount (snw) as an approximation of it. In the remaining part of the paper we refer to this TWS approximation as model TWS (mTWS). The mTWS differs in several aspects from GRACE-derived TWS: Soil moisture layers in ESMs have a depth that can vary widely between just a few and up to tens of meters depending on the model and thus does not necessarily capture the full soil moisture content at every location. Furthermore, groundwater and surface water are not explicitly included in mTWS as these states are generally not represented in CMIP5 models. However, a certain fraction of these quantities might be implicitly included in total soil moisture as the transport to ocean and atmosphere is limited and the water balance is largely closed by most of the models (Liepert & Lo, 2013). Moreover, historical CMIP5 runs do not contain regional anthropogenic intervention other than land use changes in their setup (e.g., groundwater depletion or dam building is not represented), whereas GRACE observations include their consequences. The representation of mTWS differs from model to model due to different root depths, number of soil layers, and model physics (Huang et al., 2016). Snw also exhibits large intermodel differences in representation (Brutel-Vuilmet et al., 2013). We therefore note that mTWS of different models might not be fully compatible.

After adding monthly mrso and snw for each model, we concatenate the corresponding mTWS simulations of the historical runs (1850–2005) and the RCP8.5 scenarios (2006–2100) to calculate trends for time spans that go beyond the year 2006. For those models, where more than one run is available, we calculate the ensemble mean which we regard as the most robust realization for long-term mTWS trend estimates. Afterward, the mTWS values are remapped to a common 2° × 2° geographical grid.

A bicentennial mTWS trend map (time span January 1861 to December 2099, i.e., earliest/latest common date of all models for historical/RCP8.5 experiments) is calculated from the time series of mTWS grids for each model. For each grid cell the linear trend is calculated by fitting a function

$$f(t) = a + b \cdot t + c \cdot \cos(\omega t) + d \cdot \sin(\omega t) + e \cdot \cos(2\omega t) + f \cdot \sin(2\omega t) \tag{2}$$

with parameters for bias (a), linear trend (b), annual and semiannual cycle (c, d, e, f) to the time series by means of least squares adjustment. The standard deviation of the trend is estimated from the postfit residuals. Note that we exclude the glaciated regions of Greenland, Svalbard, Gulf Coast of Alaska, and Antarctica, since not all models properly represent glacier mass balance dynamics dominating TWS in those regions.

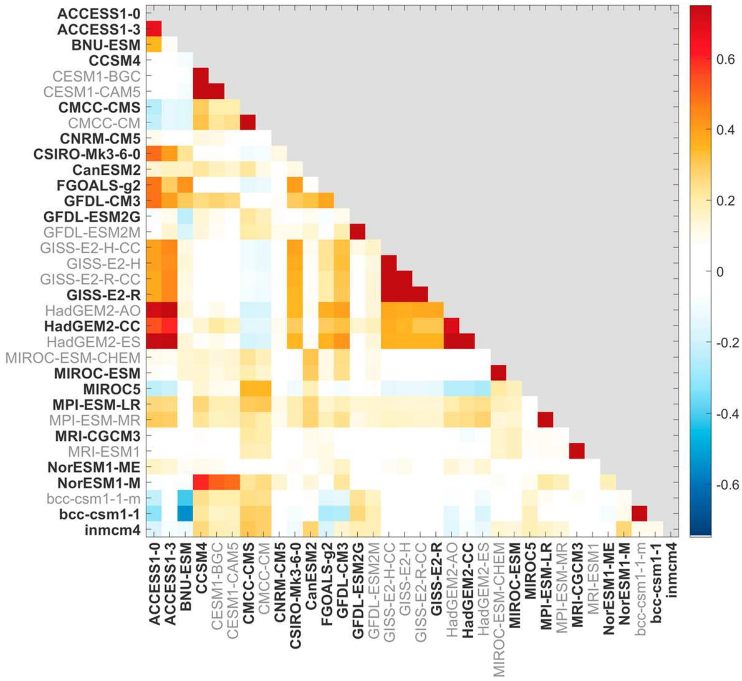


Figure 2. Correlations of the bicentennial mTWS trend maps for 34 CMIP5 models.

In total 34 CMIP5 models provide at least one run for mrso and snw. However, as some of these 34 models are either different versions of the same model or are runs with partly identical components (land surface and/or atmosphere model), it cannot be assumed that each model produces a completely independent estimate for the mrso and snw fields (Knutti et al., 2013). In order to obtain an unbiased multimodel average mTWS trend map and a reliable conclusion about model consensus, we identify the independent models by comparing the similarity of mTWS trend maps for all models. As a measure for the similarity of two maps we use the Pearson product-moment correlation coefficient r^2 calculated from the vectorized maps, giving every land pixel of the 2° × 2° grid equal weight. As trend outliers in single pixels can distort the correlation coefficient we apply a simple threshold to the mTWS trend maps, excluding absolute trend values above 23 mm/year. This is the 2σ boundary of the 14-year GRACE TWS trend (Figure 1), thus it is very unlikely that bicentennial trends above these threshold are realistic.

The correlations of the bicentennial mTWS trend maps (after applying the 23 mm/year threshold) are calculated for all 34 models and arranged in a matrix (Figure 2). Detailed information and references for the models listed in Figure 2 are given, for example, in Flato et al. (2013) and are not reiterated here. As expected, models that use common atmosphere or land surface components exhibit a very high correlation. In order to only consider models that are independent and to justify the application of equal weight to each model result, in the remaining part of the study we use only one instance from each group of models that are highly correlated ($r^2 > 75\%$). In Figure 2 the models that are excluded due to this threshold are denoted in gray font and the remaining 21 models are highlighted in bold font. The criteria for choosing a specific model among highly correlated models was based on its estimated age (most recent publication), degree of specialization (most general), or spatial resolution (closest to 2° × 2°). Generally, after excluding all but one from the highly correlated models, the correlation among trend maps from different models is very low (mean $r^2 = 10\%$, maximum $r^2 = 67\%$) and for some pairs of models it is even negative (minimum $r^2 = -55\%$). This analysis demonstrates the large inhomogeneity among CMIP5 models regarding trends in mTWS.

In order to further investigate model spread we define different time spans (Table 1) for which we calculate and discuss mTWS trend maps in the following. First, the 21 models that remain after excluding highly correlated models, are used to calculate a median trend map for the bicentennial time span January

Table 1
Notation for Different Time Spans That Are Investigated for mTWS Trends

Time span	Notation
Jan 1861 to Dec 2099	bicentennial trend
Jan 1861 to Aug 2016	historical trend
Sep 2016 to Dec 2099	RCP8.5 trend
Jan 1986 to Dec 2035	50a tendency
Jan 1996 to Dec 2025	30a tendency
Apr 2002 to Aug 2016	14a tendency

1861 to December 2099 (Figure 3a), that is, for each geographical grid cell the median of the trends of all models is determined. We use the median instead of the arithmetic unweighted mean because it is much less affected by outliers and thus can be assumed to be a more robust estimate for the trend. However, to identify nonsignificant trends in the median map (stippled regions in Figure 3a) we carry out error propagation for the arithmetic mean, because this is not straightforward for the median. Figure 3a is not affected by a model drift in mTWS, as trends from preindustrial control simulations (i.e., model runs only forced with natural, nonevolving atmospheric concentrations) of the same CMIP5 models were found to be an order of magnitude smaller and thus are negligible (not shown). According to the 21 models, the largest trends occur mainly in southern Europe and Turkey, in Central America and in the west of North America, in the north of South America and in the Himalaya region. The climatological trends derived here are in agreement with the results of a previous study (Berg et al., 2017) that focused on total soil moisture, even though significant differences are present in high latitudes since mTWS also includes snowpack.

To assess the reliability of the median mTWS trends, we compute the level of consensus of the 21 models, that is, the number of models with the same bicentennial trend direction for a given grid cell (Figure 3b). The higher the consensus, the higher the certainty that the agreement is not by chance, for example, if 15 or more of 21 models agree on the sign, the probability that this is just chance is only 4% or less (Dirmeyer et al., 2013). Hence, the higher the consensus in a grid cell, the more we can trust the direction of the trend in this grid cell according to the models. In many regions high consensus corresponds to large trends and vice versa. However, this is not valid everywhere, meaning that also the sign of small trends can be represented by a majority of models (e.g., India) and inversely, there might be model disagreement about the direction of large trends (e.g., Northern Russia). For the bicentennial mTWS trend, we find 39% of the global land area to exhibit a drying (30%) or wetting trend (9%) that is supported by at least 71% (15 of 21) of the models. These findings are not free of uncertainties as the consensus map (Figure 3b) might be affected by systematic deficits in CMIP5 models, such as in particular the lack of groundwater storage in aquifers at different depth and thus very different residence times (Pokhrel et al., 2014).

To investigate if mTWS trends as calculated for the bicentennial time span are in principle already detectable in observational records available today, we compute (in addition to the bicentennial time span) mTWS trends for a historical time span January 1861 to August 2016 (until the end of the GRACE time span; Figure 4a). For comparison, also the mTWS trends for the RCP8.5 time span September 2016 to December 2099 are displayed (Figure 4b). Overall, we find a similar pattern for the historical and the RCP8.5 trend (pattern correlation of 55%), though the historical trend has a much smaller magnitude (only about 20% of RCP8.5). Furthermore, the portion of land area where the median trends are not significant (stippled areas, 95% confidence level) is larger for the historical time span than for the RCP8.5 time span. However, in 73% of the land area the historical trend is already significant and in 68% it is in agreement with the RCP8.5 trend map. In high-consensus regions (agreement of bicentennial trend sign in ≥71% of the models, Figure 3b) to which we restrict the analysis in section 5 and 6, the area of agreement between significant historical and RCP8.5 trends is 92%. This indicates that in most regions the current trends are set to continue in the same direction and even increase in the future, thereby suggesting that the processes shaping the climate change footprint on TWS are already acting today. The consensus among the CMIP5 models is generally lower for the historical time span (Figure 4c) than for the RCP8.5 time span (Figure 4d), which is related to the fact that stronger trends generally imply higher consensus and RCP8.5 is the scenario with the strongest climate change signal. The patterns of the consensus maps are similar for all three time spans (bicentennial, historical, and RCP8.5), thus we infer that regions of large model agreement are largely independent from the selected time span (for centennial trends).

4. Influence of Observation Time Span

From Figures 3 and 4 it can be concluded that for centennial time spans a temporally stable pattern of drying and wetting trends exists in the models. However, we cannot expect to readily find these trend patterns in

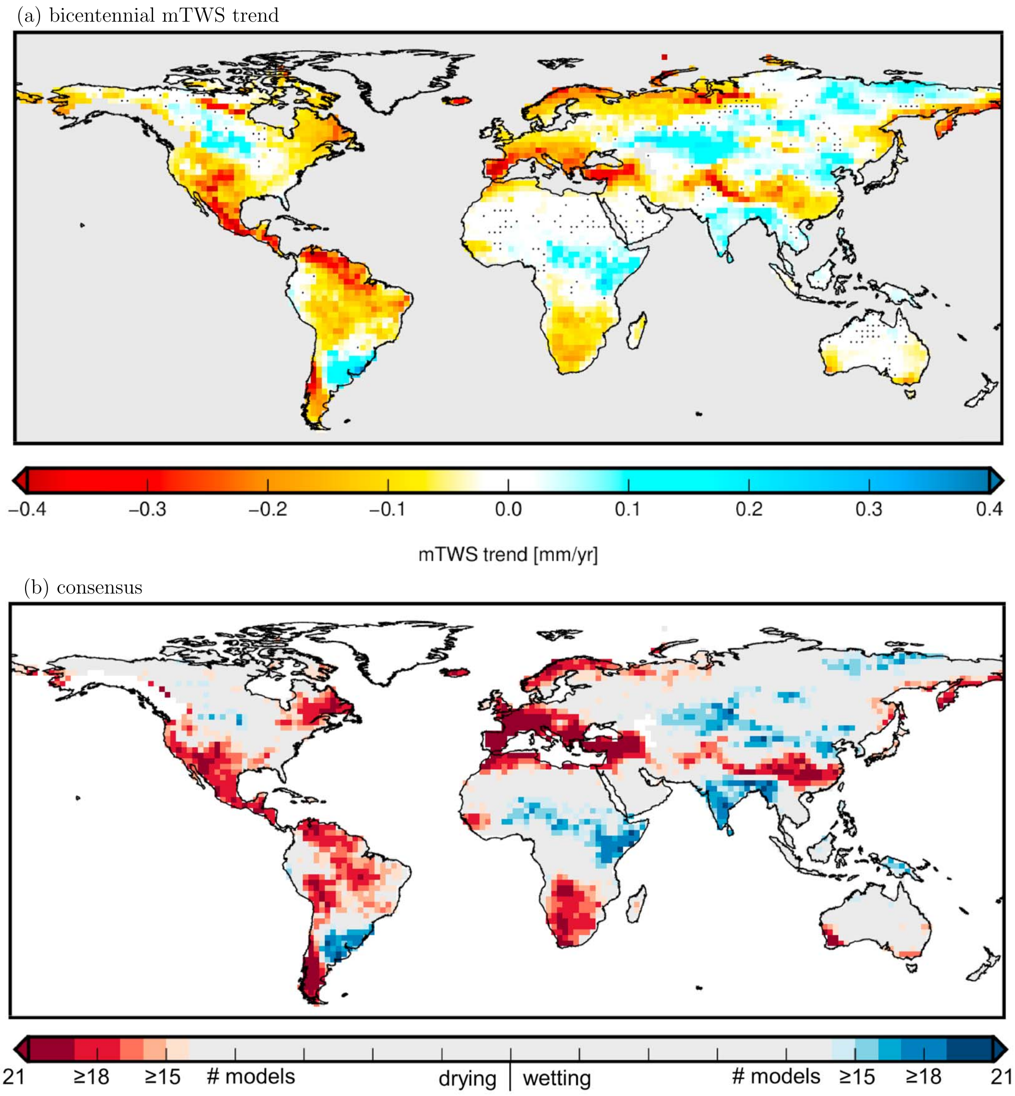


Figure 3. (a) Median of bicentennial mTWS trend maps from 21 CMIP5 models (without Greenland, Svalbard, Gulf Coast of Alaska, and Antarctica). Stippling indicates regions where the mean trend is not significantly different from zero ($\alpha = 0.05$). (b) Consensus map for bicentennial mTWS trends from 21 CMIP5 models. Red colors indicate that ≥#models agree on a negative (i.e., drying) trend, blue colors indicate that ≥#models agree on a positive (i.e., wetting) trend.

a short time period of only 14 years for which GRACE observations are available. For short time periods, interannual variations may be dominating the trend estimation in many regions of the world.

To estimate the influence of the observation time span on the expected agreement with the bicentennial trend, we perform two model studies using tendency maps for different time spans calculated from the CMIP5 models. In the first model study we investigate after which time span long-term climatic trends in mTWS might be clearly distinguished from interannual variations. In the second model study we estimate the degree to which even after long time spans the observed trends might still be in disagreement with the long-term climatic trend. In contrast to the other sections of the paper, where we rely on the ensemble means, for these model studies we only use one individual run (r1i1p1) per model in order to preserve interannual variability. This is important as natural variations would largely average out by calculating ensemble means. By using CMIP5 model output for simulating differently long observation time spans, we presume that individual model runs represent natural variability realistically in terms of relative magnitude, frequency, and duration.

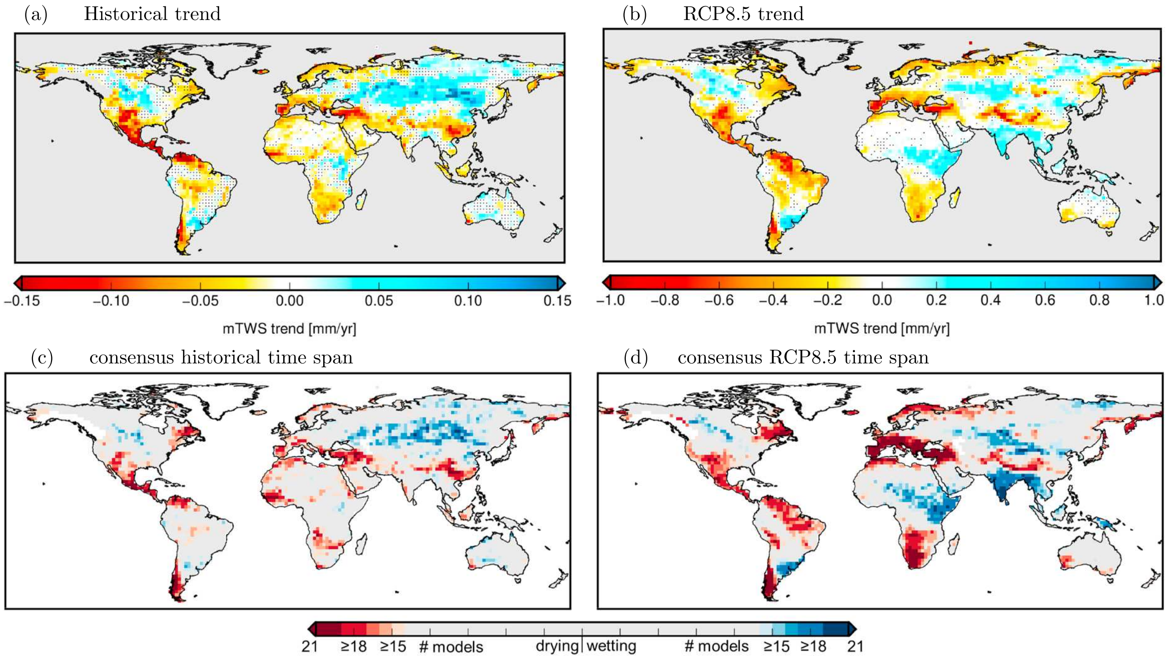


Figure 4. (top) Median of mTWS trend maps from 21 CMIP5 models for (a) historical (January 1861 to August 2016) and (b) RCP8.5 (September 2016 to December 2099) time span. Stippling indicates regions where the mean trend is not significantly different from zero ($\alpha = 0.05$). Please note the different color scales in (a) and (b). (bottom) Consensus maps for bicentennial mTWS trends from 21 CMIP5 models for (c) historical and (d) RCP8.5 time span.

For the first study we fit trends in a least squares sense for time spans of different lengths ranging from 14 to 100 years in steps of 5 years with the center year 2010. Examples for tendency maps obtained for the 50a, 30a, and 14a time periods are given in Figure 5. The median tendency maps for all 18 time spans are each correlated to the bicentennial median trend map (Figure 5a). For the 14a observation period the global correlation is only 23%, but with increasing time span it asymptotically approaches 100% (blue curve in Figure 5a). After the 30a time span the global correlation is 57% which is the same order of similarity that we find for the historical and RCP8.5 time spans (55%). Thus we conclude that around 30 years of TWS observations would be the minimum time to globally obtain a TWS trend comparable to long-term model results. However, even though the agreement between trend patterns might be low at 14a globally, this might not be the case locally, for instance, when only considering regions that exhibit strong model agreement. When calculating the correlation of the 14a tendency and the bicentennial trend only for grid cells with a model consensus of ≥71%, the correlation coefficient increases to 39% (red curve in Figure 5a), when additionally excluding nonsignificant grid cells, it increases to 52% (yellow curve in Figure 5a).

In this model study we evaluate the (global) spatial pattern correlation which only contains limited information about the agreement of trends for individual grid cells. This means that though this experiment brings out what to expect from the similarity of the spatial patterns, it does not provide the likelihood for a local mTWS tendency computed from a certain time span to actually match the bicentennial trend in that grid cell. As we are interested in regions where 14a GRACE TWS tendencies agree with bicentennial mTWS model trends and want to rate the results with respect to what to expect from this short time span, we perform a second model study: For each of the 21 CMIP5 models we cut 22 slices of 14a mTWS data with a distance of 5 years (centered around the year 1970) and estimate 22 14a tendencies. For each grid cell the fraction of tendencies that agree or disagree (in terms of sign) with the bicentennial mTWS trend from that particular model is calculated. Subsequently, the global mean of all fractions and all models is computed. This procedure is repeated for different time spans from 1 to 100 years in steps of 5 years (Figure 6). According to the models the probability that a 14a tendency is in agreement with the long-term trend is on average 53%, which is slightly better than random chance. Even after a century there is still a chance of 27% that an individual tendency does not match the bicentennial trend even though the global pattern correlation is already high with 87%. This indicates that there is natural variability in the models even over long time periods of 100 years and more, which Laepple and Huybers (2014) found to be caused by sea surface temperature

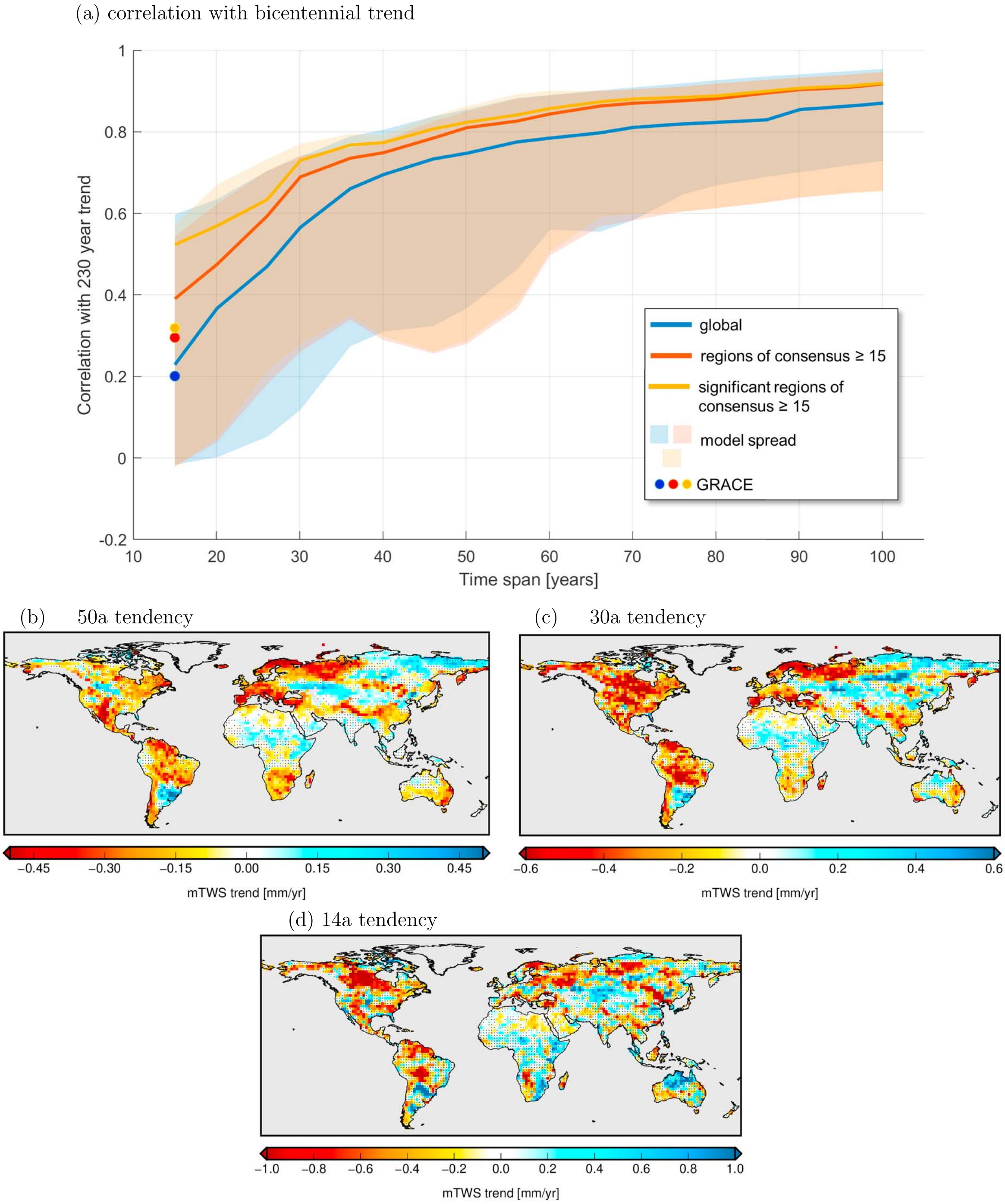


Figure 5. (a) Correlation of the mTWS tendency maps for different time spans (center year 2010) with the bicentennial mTWS trend map. (b–d) Median of mTWS tendency maps from 21 CMIP5 models for (b) 50a time span, (c) 30a time span, and (d) 14a time span. Each time span is centered around the year 2010. Note the different color scales due to larger variability for shorter time spans. Stippling indicates regions with nonsignificant trends ($\alpha = 0.05$).

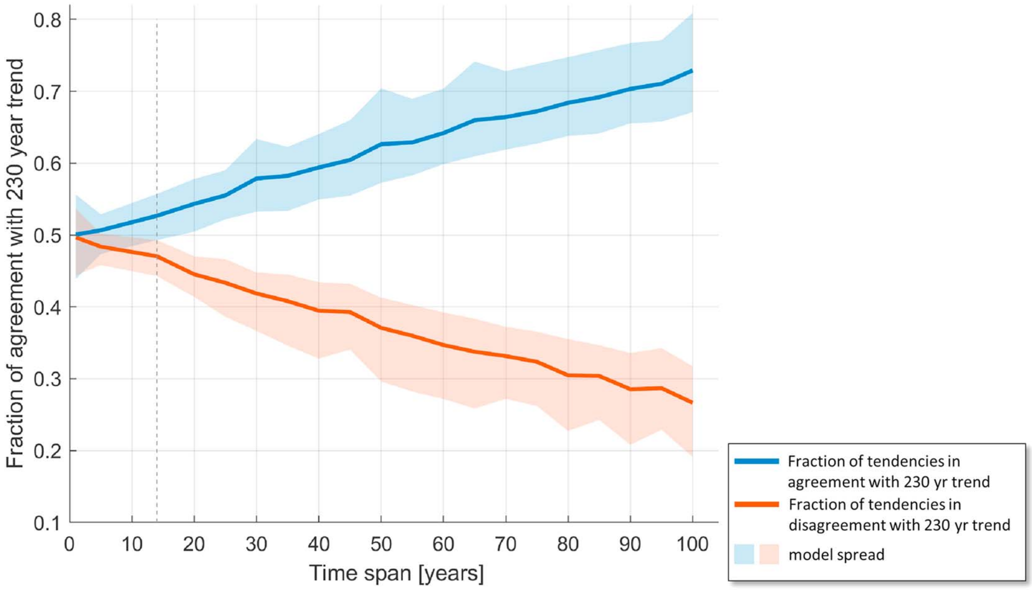


Figure 6. Proportion of grid cells from mTWS tendency maps for different time spans in agreement and disagreement with the direction of the bicentennial mTWS trend. Stippled line indicates 14 years.

variability on these time scales. However, when taking into account the model spread, we conclude that after about 14 years (stippled line in Figure 6) the two curves are just starting to be clearly distinguishable, that is, an agreement with the long-term trend can be significantly considered to being not just by chance any more. This is an indication that it is possible—at least locally—to distinguish long-term climate-driven TWS trends from interannual variations even in short observation time series and thus a comparison of 14a tendencies from GRACE to bicentennial model trends is reasonable already today.

5. Comparison of CMIP5 mTWS and GRACE TWS Trends

The two model studies in section 4 indicate that when comparing TWS tendencies from a 14a observation period to bicentennial mTWS trends we cannot expect too much agreement: a fairly low global pattern correlation and a probability of agreement only slightly above 50%. However, these model studies represent only the global mean and according to Figure 6 after 14 years we are starting to be able to clearly identify regions of long-term wetting and drying. We thus compare the TWS trends derived from GRACE observations for the 14a time period (Figure 1) to the bicentennial median mTWS trend (Figure 3a).

We note that the magnitude of the TWS trend from GRACE is substantially larger than the bicentennial median mTWS trend. One reason for this is that generally interannual TWS trends (as seen from 14 years of GRACE) exhibit a larger magnitude than bicentennial trends mostly unaffected by low-frequency climate variability. This can, for example, be seen from the trends of the median 14a tendency maps from the CMIP5 models (Figure 5d), which are already much larger than the median bicentennial trends. Individual models even exhibit a larger 14a tendency range than such maps where trend magnitudes are dampened due to the median calculation. In addition, TWS and mTWS do not represent the same physical entity everywhere as described in sections 2 and 3 and mTWS might have a lower amplitude as soil moisture depth is often limited. Recent results by Scanlon et al. (2018) show that also in off-line global hydrological models and land surface models TWS trends are generally of smaller magnitude compared to the observed GRACE trends. This indicates that even high-resolution uncoupled models driven with realistic atmospheric forcings still have difficulties to simulate realistic TWS trend magnitudes. We also note that agreement between hydrological models and GRACE could be improved by considering groundwater storage in the models (Pokhrel et al., 2013), which is entirely omitted in all CMIP5 models considered here.

As expected, the correlation between the global patterns of the 14a TWS tendency from GRACE and the bicentennial median mTWS trend is low ($r^2 = 20\%$, blue dot in Figure 5a), but it largely meets the correlation expected from the model study where we compared the 14a median mTWS tendency to the bicentennial

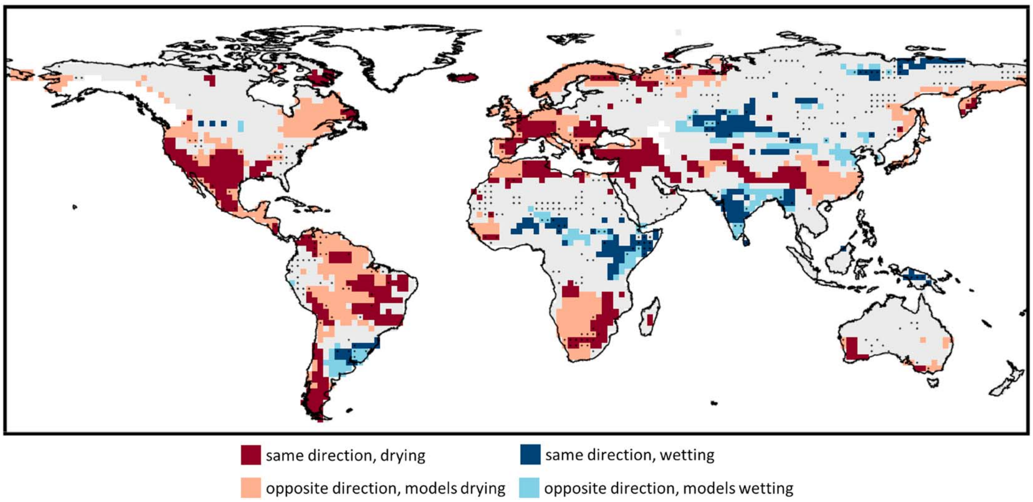

same direction, drying
same direction, wetting
opposite direction, models drying
opposite direction, models wetting

Figure 7. Regions where at least 71% of CMIP5 models show the same direction of bicentennial mTWS trend, distinguished into regions where the TWS trend from GRACE has the same or opposite direction. Stippling indicates regions with nonsignificant GRACE trends ($\alpha = 0.05$).

median mTWS trends ($r^2 = 23\%$, see section 4). When calculating the correlation only for grid cells with a model consensus of ≥71%, the correlation coefficient for GRACE increases to 30% (red dot in Figure 5a), which is slightly lower than expected from the model study ($r^2 = 39\%$) but largely within the model spread. After excluding nonsignificant GRACE trends the correlation further rises to 32%, which is, compared to the model study, in the middle of the model spread, but lower than the median ($r^2 = 52\%$). The reason for this is likely due to the different number of nonsignificant trend grid cells for GRACE and for the model median (compare Figure 1 and Figure 5d).

In Figure 7 areas are displayed in red where at least 71% (15 of 21) of the models show a negative bicentennial trend direction (Figure 3). A grid cell is marked in dark red if at the same location the trend from GRACE (Figure 1) exhibits a negative sign as well, or in light red if the trend from GRACE has the opposite (i.e., positive) sign. Analogously, positive high-consensus model trends are marked in blue; dark blue where GRACE has the same (positive) sign and light blue where GRACE has the opposite (negative) sign. Dark red or dark blue areas in Figure 7 indicate hot spot regions where trends in GRACE data may already be related to climate change signals because the majority of CMIP5 models supports this trend (at least regarding its sign) on the bicentennial time scale. According to Figure 6, hot spot regions of drying trends are mainly the region around the Mediterranean Sea, southwestern United States, the southern tip of South America, and the Himalaya region.

Few wetting trends are identified as hot spots, potentially in South America, Central Africa, Central and North Asia, and India. This is quite consistent with the bicentennial mTWS trend map (Figure 3a), which is dominated by negative trends (66% of trends are negative, 34% positive, global mean -0.04 mm/year, excluding Greenland, Alaska, and Antarctica). A reason for dominating drying trends in CMIP5 models might be that the models have a limited ability to capture anomalously high water storage. In most models that do not include groundwater, rivers, lakes, and wetlands, water that exceeds the soil moisture storage capacity is allocated to surface runoff and does not reenter the land surface scheme. In addition, an underestimation of persistence times of water in the soil might prevent the simulation of high accumulations of water. Thus, models tend to simulate drying trends more easily than wetting trends.

Besides the dark red and blue areas there are several regions of opposite signs between 14a GRACE TWS tendencies and bicentennial mTWS model trends (light areas in Figure 7). Major regions of disagreement are located in northeastern Canada, Fennoscandia, southeastern China, southern Africa, and central South America. The proportion of high-consensus areas agreeing with GRACE versus those disagreeing with GRACE is 49% to 51%. When not restricting to high-consensus regions but calculating the proportion globally it is 50% versus 50%. From the second model study in section 4 we obtained a higher percentage of agreement of about 53%, but this number is estimated by comparing each 14a model tendency map to the

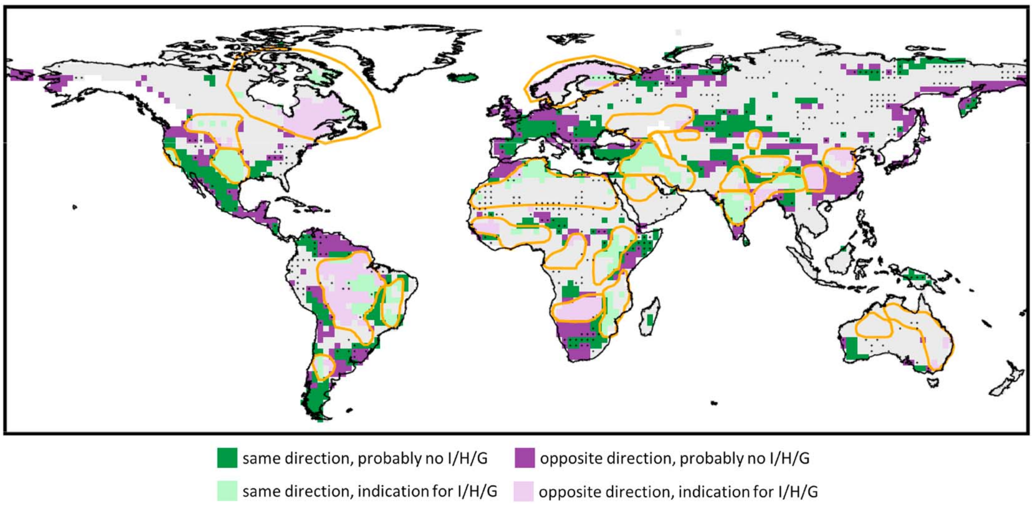

same direction, probably no I/H/G
opposite direction, probably no I/H/G
same direction, indication for I/H/G
opposite direction, indication for I/H/G

Figure 8. Regions where at least 71% of CMIP5 models show the same direction of bicentennial mTWS trend, distinguished into regions where the TWS trend from GRACE has the same (green color) or opposite (violet color) direction. Regions affected by interannual variability (I), human impact (H), or glacial isostatic adjustment (G) are outlined by orange polygons and shaded in light green/violet. Stippling indicates regions with nonsignificant GRACE trends ($\alpha = 0.05$).

bicentennial trend map from the same model. As we do not have a bicentennial reference for GRACE but compare to the model median instead, it is expected that the match is not at the same level. Furthermore, results from the model study might be too optimistic if modeled interannual variability is lower than in reality. In addition, the results from section 4 are not directly comparable to GRACE as human impact such as dam building and groundwater abstraction is not reflected in the model median which further affects the level of agreement. From the low level of agreement between modeled and observed TWS trends we conclude that in many regions the influence of interannual variations in the 14a time span is still dominant at this stage and thus the comparison to long-term modeled trends is affected by large uncertainties. However, the analysis of the existing hot spot and noncompliance regions in Figure 7 can give valuable information on potential climate-related wetting and drying as well as indications for possible model shortcomings and can be the focus of future investigations.

6. Uncertainty Analysis

Due to interannual variability in the short GRACE observation time span and human impacts that are not considered by the models, we cannot assume the same level of certainty for the results presented in section 5 in every region. In order to identify regions where climate-related wetting or drying trends may be overlaid by interannual variability (I) or human impact (H) we access a study by Rodell et al. (2018), who attribute the TWS trends from GRACE for April 2002 to March 2016 to their different dominating origins for 34 major basins. The regions where I or H is present in GRACE TWS trends according to Rodell et al. (2018) are outlined as orange polygons in Figure 8. Within these orange polygons the results from Figure 7 about agreement/disagreement have to be regarded as uncertain due to I/H effects. In addition, we marked regions affected by glacial isostatic adjustment (GIA; G) as regions possibly not comparable with mTWS from ESMs as residual GIA effects in GRACE might remain in the trend even after removing a state-of-the-art GIA model. There might be other regions affected by I/H/G that are wrongly not considered as uncertain because they are not accounted for by Rodell et al. (2018) who focused on the 34 study regions with the most prominent trend signals in GRACE. Furthermore, there might be additional regions where mTWS trends from CMIP5 models are systematically wrong—for example, due to missing groundwater and deeper soil layers, but these regions are also not explicitly identified and marked as uncertain here.

By means of the orange polygons in Figure 8 we can distinguish the hot spot and noncompliance regions (see Figure 7) into four categories:

1. Same direction of model majority and GRACE trend, and no indication for I/H/G (dark green). In these regions we assume the climate-related wetting or drying simulated by the models to be confirmed by observations.
2. Same direction, but indication for I/H/G (light green). Here we do not know at this time if a possible underlying climatic trend has the same or opposite direction as modeled trends, thus we mark these regions as uncertain. By extending the observation record it might turn out in the future that the current match in some of these regions is only chance due to I/H/G effects counteracting the long-term trend.
3. Opposite direction, and no indication for I/H/G (dark violet). In these regions we have no reason to assume that the direction of the 14a tendency from observations differs from the real long-term climate trend, thus there might be deficits in CMIP5 models leading to a mismatch. Of course, also other I/H/G effects that were not yet identified might be the reason for noncompliance.
4. Opposite direction, but indication for I/H/G (light violet). As in Category 2, we do not know the direction of a possible underlying climatic trend, thus consider these regions as uncertain. The current mismatch might be only chance due to I/H/G effects dominating the 14a GRACE trends and in the long term these regions might turn out to be agreement regions.

In Category 1, where observations already today confirm what a majority of models predict for the long-term, mainly three larger regions remain: drying conditions in southwestern United States/Mexico and around the Mediterranean Sea (central southern Europe and Turkey), and wetting conditions in Central Asia (indicated by a cluster of dark green grid cells). These regions we regard as hot spot regions of drying and wetting that can already today be attributed to climate change with the help of GRACE.

Category 2 indicates regions where potential climate signals are overlaid by I/H/G effects that might be the cause for the current agreement. However, in some regions it is quite likely that the long-term trend has the same direction as the I/H/G effects, for example, the light green spots in central southern United States and the Middle East (i.e., Syria, Iraq, and Iran), where relatively large connected regions of agreement are adjacent to Category 1 regions. According to Rodell et al. (2018) in both regions the large negative TWS trends seem to be due to a combination of drought and enhanced groundwater depletion. Additionally, in the Middle East dam building in Turkey further intensifies the lack of water (Voss et al., 2013). Also at parts of the Mediterranean coast in northern Africa models and GRACE see a common drying that might be connected to the long-term drying conditions in southern Europe, even though the currently observed trend is assumed to be dominated by groundwater abstraction (Döll et al., 2014). Another location where I/H/G effects might enhance a common long-term trend is the light green spot in India. Here, possibly climate-related precipitation increase is overlaid by human impact in form of a groundwater policy change that is reflected in overall wetting conditions (Bhanja et al., 2017).

In Category 3, a mismatch of model majority and GRACE is unlikely to be due to I/H/G, thus these regions indicate possible systematical errors in CMIP5 models. For example, CMIP5 models were found to underestimate summer precipitation over southeastern China (Chen & Frauenfeld, 2014), which might explain the negative soil moisture trend not supported by GRACE. Furthermore, systematic rainfall biases in CMIP5 models over southern Africa were identified by Munday and Washington (2018), another possible reason for noncompliance with GRACE. However, this needs to be further investigated, as according to Munday and Washington (2018) rainfall is largely overestimated whereas soil moisture is simulated to decrease. Large connected regions of noncompliance in western and northeastern Europe are likely affected by interannual variations. As the apparent positive TWS trends seen by GRACE are very small here (see Figure 1), hardly significant, and dominated by the annual cycle they were not accounted for by Rodell et al. (2018). The same holds for a dark violet region in northern South America.

Category 4 denotes regions affected by I/H/G that are not in agreement with long-term predicted trends. We consider these regions as uncertain as possible climate signals masked by I/H/G might actually agree with model trends. Large Category 4 areas are the GIA-affected regions Fennoscandia and northeastern Canada where the uncertainty from the applied GIA model might be dominating the GRACE trend. In the Amazon region in central South America a recovery from a drought in the early GRACE period is responsible for an overall wetting trend in GRACE. However, the mismatch could also result from model deficits due to missing groundwater. Considering groundwater buffering effects in CMIP5 models causes a shift in the evapotranspiration regime resulting in much less drying trends in the Amazon region (Pokhrel et al., 2014). In southern Africa the Category 4 region is associated with interannual variability, a progression from dry to wet during the GRACE period. The light violet spot in southeastern China can be connected to the Three Gorges Dam reservoir filling that models do not capture. However, it is not clear to what extent also model deficits in southeastern China and southern Africa are responsible for this mismatch as these are adjacent to Category 3 regions.

7. Conclusions

We analyzed in this study long-term trends in TWS derived from a selection of 21 coupled climate models stored in the CMIP5 archive and compared them to satellite observed TWS trends from 14 years (April 2002 to August 2016) of GRACE data. We found a large disagreement among the bicentennial (January 1861 to December 2099) TWS trends (sum of mrso and snw) from different models: the mean correlation among individual models is only 10%, which reflects the still high uncertainties in TWS variability simulated by present-day global coupled climate models. While a significant part of intermodel variation might result from the different atmosphere components, differences in land surface parameterization, particularly the soil moisture and snow storage capacities, are likely to add to this. We nevertheless identified several regions of high model consensus regarding the direction of the trend on which we focused for the comparison with GRACE-derived TWS. Furthermore, we found a large agreement between modeled TWS trends for the historical (January 1861 to August 2016) and the RCP8.5 (September 2016 to December 2099) time span, indicating that long-term TWS trends are already emerging in present time and thus can be expected to be contained in observational records available today.

By computing model TWS tendencies for differently long time spans (from 14 to 100 years) and comparing them to the bicentennial trend map, we assessed the influence of interannual variations on the degree of agreement between long-term and short-term trends. For the global pattern correlation we concluded that a time span of 30 years or more would be sufficient to globally distinguish interannual climate variability in TWS from long-term climate trends. However, for even 14 years we obtained a global correlation of 23% with the bicentennial TWS trend, which regionally is substantially higher, for instance, when limiting to significant trends in regions of strong model consensus only (52%). When estimating the fraction of grid cells with the same direction of the trend for the bicentennial and different shorter time spans we found that after 14 years the proportion is 53% agreement versus 47% disagreement, and even after a century there is still a substantial probability of disagreement (27%). We identified 14 years as the minimum observation time span required to distinguish long-term trends from interannual variations, with a (global) probability that is better than just chance.

By comparing the bicentennial modeled median TWS trend map against the TWS trend map obtained from GRACE data over the period April 2002 to August 2016, we found a similar global correlation (20%) as we expected from the model study. Focusing on regions of strong model consensus, we identified hot spot regions where modeled TWS trends have the same direction as the GRACE TWS trend. Drying hot spots were mainly found in the region around the Mediterranean Sea, the west coast of North America, the southern tip of South America, and the Himalaya region. Wetting hot spots are sparse and only found for small areas in South America, Central Africa, Central and North Asia, and India. The largest regions of noncompliance between trends from models and GRACE are located in northeastern Canada, Fennoscandia, southeastern China, southern Africa, and central South America. In total, the proportion of regions in agreement versus regions in disagreement in high model consensus regions is 49% versus 51%, which indicates that the 14-year GRACE time series is in many regions still dominated by interannual variations.

We further investigated the existing hot spot and noncompliance regions regarding possible natural interannual variability or human impact in order to classify the results regionally. In this classification, 36% of the high-consensus area was marked as uncertain and thus remains undetermined at this time if and how much it is affected by climate change. In the other 64% of the high-consensus area the climate signal was assumed to dominate the TWS trends. Within these 64% the proportion of agreement versus disagreement areas is 49% versus 51%, which is the same proportion as in the entire high-consensus area without uncertainty assessment. From this we conclude, that in half of the area marked as certain either model deficits or unidentified interannual variability in observations cause disagreement. This is consistent with our findings from the model study that revealed a large influence of interannual variations on TWS trends even after long time spans.

Nevertheless, from the GRACE record available today, the classification leads to the identification of hot spot regions in southern United States/Mexico, central southern Europe, and central Asia, where GRACE confirms modeled long-term drying or wetting trends already today. In some regions, for example, in southeastern China, the Amazon region, and in southern Africa the results hint at possible model deficits, for example, due to missing groundwater modeling. Furthermore, we identified regions where mismatch of models and observations might be due to interannual variations or other effects (e.g., dam building, groundwater withdrawal, glacial isostatic adjustment) in GRACE that are not included in the CMIP5 models.

As the time series of GRACE data will soon be extended with new observations from the GRACE-FO mission launched in May 2018, observed TWS trends are expected to become even more representative for the long-term climate response. At the same time, climate model outputs from the upcoming CMIP6 will include new model versions with altered representations of soil moisture-related variables, which might either improve or degrade the model consensus depending on the impact of new parameterizations and the effect of a higher degree of model freedom and complexity. Comparisons between observed and modeled TWS trends as presented in this paper will thus remain of high relevance for future climate model assessments.

Acknowledgments
We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Figure 2 of this paper) for producing and making available their model output (e.g., at "https://esgf-data.dkrz.de/search/cmip5-dkrz/"). For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. We are grateful to Torsten Mayer-Gürr and his group at TU Graz, Austria, for providing the ITSG-Grace2018s data at this website (https://www.tugraz.at/institute/ifg/downloads/gravity-field-models/itsg-grace2018/). The degrees 1 and 2 harmonic coefficients used during GRACE processing were received from this site (http://grace.jpl.nasa.gov).

References

A, G., Wahr, J., & Zhong, S. (2013). Computations of the viscoelastic response of a 3-D compressible Earth to surface loading: An application to Glacial Isostatic Adjustment in Antarctica and Canada. *Geophysical Journal International*, *192*(2), 557–572. https://doi.org/10.1093/gji/ggs030

Berg, A., Sheffield, J., & Milly, P. C. D. (2017). Divergent surface and total soil moisture projections under global warming. *Geophysical Research Letters*, *44*, 236–244. https://doi.org/10.1002/2016GL071921

Bhanja, S. N., Mukherjee, A., Rodell, M., Wada, Y., Chattopadhyay, S., Velicogna, I., et al. (2017). Groundwater rejuvenation in parts of India influenced by water-policy change implementation. *Scientific Reports*, *7*(1), 7453. https://doi.org/10.1038/s41598-017-07058-2

Bhattacharyya, S., & Narasimha, R. (2005). Possible association between Indian monsoon rainfall and solar activity. *Geophysical Research Letters*, *32*, L05813. https://doi.org/10.1029/2004GL021044

Brutel-Vuilmet, C., Ménégoz, M., & Krinner, G. (2013). An analysis of present and future seasonal Northern Hemisphere land snow cover simulated by CMIP5 coupled climate models. *The Cryosphere*, *7*(1), 67–80. https://doi.org/10.5194/tc-7-67-2013

Caron, L., Ivins, E. R., Larour, E., Adhikari, S., Nilsson, J., & Blewitt, G. (2018). GIA model statistics for GRACE hydrology, cryosphere, and ocean science. *Geophysical Research Letters*, *45*, 2203–2212. https://doi.org/10.1002/2017GL076644

Chen, L., & Frauenfeld, O. W. (2014). A comprehensive evaluation of precipitation simulations over China based on CMIP5 multimodel ensemble projections. *Journal of Geophysical Research: Atmospheres*, *119*, 5767–5786. https://doi.org/10.1002/2013JD021190

Cheng, M., Tapley, B. D., & Ries, J. C. (2013). Deceleration in the Earth's oblateness. *Journal of Geophysical Research: Solid Earth*, *118*, 740–747. https://doi.org/10.1002/jgrb.50058

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., et al. (2013). Long-term climate change: Projections, commitments and irreversibility. In T. F. Stocker (Ed.), *Climate change 2013: The physical science basis. Contribution of working group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 1029–1136). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.024

Committee on the Decadal Survey for Earth Science and Applications from Space, Space Studies Board, Division on Engineering and Physical Sciences, & National Academies of Sciences, Engineering, and Medicine (2018). *Thriving on our changing planet: A decadal strategy for earth observation from space*. Washington, DC: National Academies Press. https://doi.org/10.17226/24938

Dirmeyer, P. A., Jin, Y., Singh, B., & Yan, X. (2013). Trends in land-atmosphere interactions from CMIP5 simulations. *Journal of Hydrometeorology*, *14*(3), 829–849. https://doi.org/10.1175/JHM-D-12-0107.1

Döll, P., Schmied, H. M., Schuh, C., Portmann, F. T., & Eicker, A. (2014). Global-scale assessment of groundwater depletion and related groundwater abstractions: Combining hydrological modeling with information from well observations and GRACE satellites. *Water Resources Research*, *50*, 5698–5720. https://doi.org/10.1002/2014WR015595

Dorigo, W. A., Gruber, A., De Jeu, R. A. M., Wagner, W., Stacke, T., Loew, A., et al. (2015). Evaluation of the ESA CCI soil moisture product using ground-based observations. *Remote Sensing of Environment*, *162*, 380–395. https://doi.org/10.1016/j.rse.2014.07.023

Eicker, A., Forootan, E., Springer, A., Longuevergne, L., & Kusche, J. (2016). Does GRACE see the terrestrial water cycle ”intensifying”? *Journal of Geophysical Research: Atmospheres*, *121*, 733–745. https://doi.org/10.1002/2015JD023808

Eicker, A., Schumacher, M., Kusche, J., Döll, P., & Schmied, H. M. (2014). Calibration/data assimilation approach for integrating GRACE data into the WaterGAP Global Hydrology Model (WGHM) using an ensemble Kalman filter: First results. *Surveys in Geophysics*, *35*, 1285–1309. https://doi.org/10.1007/s10712-014-9309-8

Fagiolini, E., Flechtner, F., Horwath, M., & Dobslaw, H. (2015). Correction of inconsistencies in ECMWF's operational analysis data during de-aliasing of GRACE gravity models. *Geophysical Journal International*, *202*(3), 2150–2158. https://doi.org/10.1093/gji/ggv276

Famiglietti, J. S., & Rodell, M. (2013). Water in the balance. *Science*, *340*(6138), 1300–1301. https://doi.org/10.1126/science.1236460

Fasullo, J. T., Lawrence, D. M., & Swenson, S. C. (2016). Are GRACE-era terrestrial water trends driven by anthropogenic climate change? *Advances in Meteorology*, *2016*, 4830603. https://doi.org/10.1155/2016/4830603

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., et al. (2013). Evaluation of climate models, *Climate change 2013: The physical science basis. Contribution of working group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 741–882). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781107415324.020

Flechtner, F., Neumayer, K.-H., Dahle, C., Dobslaw, H., Fagiolini, E., Raimondo, J.-C., & Güntner, A. (2016). What can be expected from the GRACE-FO laser ranging interferometer for earth science applications? *Surveys in Geophysics*, *37*(2), 453–470. https://doi.org/10.1007/s10712-015-9338-y

Freedman, F. R., Pitts, K. L., & Bridger, A. F. C. (2014). Evaluation of CMIP climate model hydrological output for the Mississippi River basin using GRACE satellite observations. *Journal of Hydrology*, *519*, 3566–3577. https://doi.org/10.1016/j.jhydrol.2014.10.036

Gray, L. J., Anstey, J. A., Kawatani, Y., Lu, H., Osprey, S., & Schenzinger, V. (2018). Surface impacts of the Quasi Biennial Oscillation. *Atmospheric Chemistry and Physics*, *18*(11), 8227–8247. https://doi.org/10.5194/acp-18-8227-2018

Güntner, A. (2008). Improvement of global hydrological models using GRACE data. *Surveys in Geophysics*, *29*(4), 375–397. https://doi.org/10.1007/s10712-008-9038-y

Guo, Z., & Dirmeyer, P. A. (2006). Evaluation of the second global soil wetness project soil moisture simulations: 1. Intermodel comparison. *Journal of Geophysical Research*, *111*, D22S02. https://doi.org/10.1029/2006JD007233

Han, S.-C., Sauber, J., & Luthcke, S. (2010). Regional gravity decrease after the 2010 Maule (Chile) earthquake indicates large-scale mass redistribution. *Geophysical Research Letters*, *37*, L23307. https://doi.org/10.1029/2010GL045449

Han, S.-C., Sauber, J., Luthcke, S. B., Ji, C., & Pollitz, F. F. (2008). Implications of postseismic gravity change following the great 2004 Sumatra-Andaman earthquake from the regional harmonic analysis of GRACE intersatellite tracking data. *Journal of Geophysical Research*, *113*, B11413. https://doi.org/10.1029/2008JB005705

Hirschi, M., & Seneviratne, S. I. (2017). Basin-scale water-balance dataset (BSWB): An update. *Earth System Science Data*, *9*(1), 251–258. https://doi.org/10.5194/essd-9-251-2017

Huang, Y., Gerber, S., Huang, T., & Lichstein, J. W. (2016). Evaluating the drought response of CMIP5 models using global gross primary productivity, leaf area, precipitation, and soil moisture data. *Global Biogeochemical Cycles*, *30*, 1827–1846. https://doi.org/10.1002/2016GB005480

Iles, C. E., Hegerl, G. C., Schurer, A. P., & Zhang, X. (2013). The effect of volcanic eruptions on global precipitation. *Journal of Geophysical Research: Atmospheres*, *118*, 8770–8786. https://doi.org/10.1002/jgrd.50678

Knutti, R., Masson, D., & Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research Letters*, *40*, 1194–1199. https://doi.org/10.1002/grl.50256

Kusche, J. (2007). Approximate decorrelation and non-isotropic smoothing of time-variable GRACE-type gravity field models. *Journal of Geodesy*, *81*(11), 733–749. https://doi.org/10.1007/s00190-007-0143-3

Laepple, T., & Huybers, P. (2014). Ocean surface temperature variability: Large model-data differences at decadal and longer periods. *Proceedings of the National Academy of Sciences*, *111*(47), 16,682–16,687. https://doi.org/10.1073/pnas.1412077111

Lambeck, K. (1988). *Geophysical geodesy: The slow deformations of the earth*. Oxford [Oxfordshire], New York: Clarendon Press: Oxford University Press. https://trove.nla.gov.au/version/21392411

Liepert, B. G., & Lo, F. (2013). CMIP5 update of 'Inter-model variability and biases of the global water cycle in CMIP3 coupled climate models'. *Environmental Research Letters*, *8*(2), 29401. https://doi.org/10.1088/1748-9326/8/2/029401

Mayer-Gürr, T., Behzadpour, S., Ellmer, M., Kvas, A., Klinger, B., & Zehentner, N. (2018). ITSG-Grace2018—Monthly, daily and static gravity field solutions from GRACE. GFZ Data Services, https://doi.org/10.5880/ICGEM.2018.003

Munday, C., & Washington, R. (2018). Systematic climate model rainfall biases over Southern Africa: Links to moisture circulation and topography. *Journal of Climate*, *31*(18), 7533–7548. https://doi.org/10.1175/JCLI-D-18-0008.1

Ni, S., Chen, J., Wilson, C. R., Li, J., Hu, X., & Fu, R. (2018). Global terrestrial water storage changes and connections to ENSO events. *Surveys in Geophysics*, *39*(1), 1–22. https://doi.org/10.1007/s10712-017-9421-7

Phillips, T., Nerem, R. S., Fox-Kemper, B., Famiglietti, J. S., & Rajagopalan, B. (2012). The influence of ENSO on global terrestrial water storage using GRACE. *Geophysical Research Letters*, *39*, L16705. https://doi.org/10.1029/2012GL052495

Pokhrel, Y. N., Fan, Y., & Miguez-Macho, G. (2014). Potential hydrologic changes in the Amazon by the end of the 21st century and the groundwater buffer. *Environmental Research Letters*, *9*(8), 84004. https://doi.org/10.1088/1748-9326/9/8/084004

Pokhrel, Y. N., Fan, Y., Miguez-Macho, G., Yeh, PatJ.-F., & Han, S.-C. (2013). The role of groundwater in the Amazon water cycle: 3. Influence on terrestrial water storage computations and comparison with GRACE. *Journal of Geophysical Research: Atmospheres*, *118*, 3233–3244. https://doi.org/10.1002/jgrd.50335

Rodell, M., Famiglietti, J. S., Chen, J., Seneviratne, S. I., Viterbo, P., Holl, S., & Wilson, C. R. (2004). Basin scale estimates of evapotranspiration using GRACE and other observations. *Geophysical Research Letters*, *31*, L20504. https://doi.org/10.1029/2004GL020873

Rodell, M., Famiglietti, J. S., Wiese, D. N., Reager, J. T., Beaudoing, H. K., Landerer, F. W., & Lo, M.-H. (2018). Emerging trends in global freshwater availability. *Nature*, *557*(7707), 651. https://doi.org/10.1038/s41586-018-0123-1

Scanlon, B. R., Zhang, Z., Save, H., Sun, A. Y., Müller Schmied, H., van Beek, L. P. H., et al. (2018). Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. *Proceedings of the National Academy of Sciences*, *115*(6), E1080–E1089. https://doi.org/10.1073/pnas.1704665115

Swenson, S., Chambers, D., & Wahr, J. (2008). Estimating geocenter variations from a combination of GRACE and ocean model output. *Journal of Geophysical Research*, *113*, B08410. https://doi.org/10.1029/2007JB005338

Syed, T. H., Famiglietti, J. S., Rodell, M., Chen, J., & Wilson, C. R. (2008). Analysis of terrestrial water storage changes from GRACE and GLDAS. *Water Resources Research*, *44*, W02433. https://doi.org/10.1029/2006WR005779

Tapley, B. D., Bettadpur, S., Watkins, M., & Reigber, C. (2004). The gravity recovery and climate experiment: Mission overview and early results. *Geophysical Research Letters*, *31*, L09607. https://doi.org/10.1029/2004GL019920

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2011). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

Trenberth, K. E. (Ed.) (2010). *Climate system modeling*. Cambridge: Cambridge University Press. OCLC: 845631695.

Voss, K. A., Famiglietti, J. S., Lo, M., Linage, C. d, Rodell, M., & Swenson, S. C. (2013). Groundwater depletion in the Middle East from GRACE with implications for transboundary water management in the Tigris-Euphrates-Western Iran region. *Water Resources Research*, *49*, 904–914. https://doi.org/10.1002/wrcr.20078

Wartenburger, R., Seneviratne, S. I., Hirschi, M., Chang, J., Ciais, P., Deryng, D., et al. (2018). Evapotranspiration simulations in ISIMIP2a—Evaluation of spatio-temporal characteristics with a comprehensive ensemble of independent datasets. *Environmental Research Letters*, *13*, 75001. https://doi.org/10.1088/1748-9326/aac4bb

Yuan, S., & Quiring, S. M. (2017). Evaluation of soil moisture in CMIP5 simulations over the contiguous United States using in situ and satellite observations. *Hydrology and Earth System Sciences*, *21*(4), 2203–2218. https://doi.org/10.5194/hess-21-2203-2017

Zhang, L., Dobslaw, H., Stacke, T., Güntner, A., Dill, R., & Thomas, M. (2017). Validation of terrestrial water storage variations as simulated by different global numerical models with GRACE satellite observations. *Hydrology and Earth System Sciences*, *21*(2), 821–837. https://doi.org/10.5194/hess-21-821-2017

A.2. Emerging Changes in Terrestrial Water Storage Variability as a Target for Future Satellite Gravity Missions

Reference

Jensen, L., Eicker, A., Dobslaw, H., Pail, R. (2020). Emerging Changes in Terrestrial Water Storage Variability as a Target for Future Satellite Gravity Missions. Remote Sensing 12, 3898. https://doi.org/10.3390/rs12233898

Abstract

Climate change will affect the terrestrial water cycle during the next decades by impacting the seasonal cycle, interannual variations, and long-term linear trends of water stored at or beyond the surface. Since 2002, terrestrial water storage (TWS) has been globally observed by the Gravity Recovery and Climate Experiment (GRACE) and its follow-on mission (GRACE-FO). Next Generation Gravity Missions (NGGMs) are planned to extend this record in the near future. Based on a multi-model ensemble of climate model output provided by the Coupled Model Intercomparison Project Phase 6 (CMIP6) covering the years 2002 – 2100, we assess possible changes in TWS variability with respect to present-day conditions to help defining scientific requirements for NGGMs. We find that present-day GRACE accuracies are sufficient to detect amplitude and phase changes in the seasonal cycle in a third of the land surface, whereas a five times more accurate double-pair mission could resolve such changes almost everywhere outside the most arid landscapes of our planet. We also select one individual model experiment out of the CMIP6 ensemble that closely matches both GRACE observations and the multi-model median of all CMIP6 realizations, which might serve as basis for satellite mission performance studies extending over many decades to demonstrate the suitability of NGGM satellite missions to monitor long-term climate variations in the terrestrial water cycle.

Declaration of own contribution

Table A.2.: Contribution to Paper No. 2.

Involved in	Estimated contribution
Ideas and conceptual design	90%
Computation and results	100%
Analysis and interpretation	90%
Manuscript, figures and tables	90%
Total	92,5%

Confirmation of Co-Authors

I hereby confirm the correctness of the declaration of the contribution of Laura Jensen for Paper No. 2 in Table A.2:

26.05.21

Annette Eicker
(HCU Hamburg)
Date

Henryk Dobslaw
Digital unterschrieben von Henryk Dobslaw
Datum: 2021.05.26 20:27:42 +02'00'

Henryk Dobslaw
(GFZ Potsdam)
Date

Roland Pail
28.05.2021

Roland Pail
(TU München)
Date

remote sensing

MDPI

Article

Emerging Changes in Terrestrial Water Storage Variability as a Target for Future Satellite Gravity Missions

Laura Jensen [1,*], Annette Eicker [1], Henryk Dobslaw [2] and Roland Pail [3]

[1] Geodesy and Geoinformatics, HafenCity University, 20457 Hamburg, Germany; annette.eicker@hcu-hamburg.de

[2] Helmholtz Centre Potsdam, German Research Centre for Geosciences (GFZ), 14473 Potsdam, Germany; dobslaw@gfz-potsdam.de

[3] Institute of Astronomical and Physical Geodesy, Technische Universität München, 80333 München, Germany; roland.pail@tum.de

* Correspondence: laura.jensen@hcu-hamburg.de

Received: 13 October 2020; Accepted: 23 November 2020; Published: 27 November 2020

check for updates

Abstract: Climate change will affect the terrestrial water cycle during the next decades by impacting the seasonal cycle, interannual variations, and long-term linear trends of water stored at or beyond the surface. Since 2002, terrestrial water storage (TWS) has been globally observed by the Gravity Recovery and Climate Experiment (GRACE) and its follow-on mission (GRACE-FO). Next Generation Gravity Missions (NGGMs) are planned to extend this record in the near future. Based on a multi-model ensemble of climate model output provided by the Coupled Model Intercomparison Project Phase 6 (CMIP6) covering the years 2002–2100, we assess possible changes in TWS variability with respect to present-day conditions to help defining scientific requirements for NGGMs. We find that present-day GRACE accuracies are sufficient to detect amplitude and phase changes in the seasonal cycle in a third of the land surface, whereas a five times more accurate double-pair mission could resolve such changes almost everywhere outside the most arid landscapes of our planet. We also select one individual model experiment out of the CMIP6 ensemble that closely matches both GRACE observations and the multi-model median of all CMIP6 realizations, which might serve as basis for satellite mission performance studies extending over many decades to demonstrate the suitability of NGGM satellite missions to monitor long-term climate variations in the terrestrial water cycle.

Keywords: terrestrial water storage; GRACE; CMIP6; climate models; climate projections; variability; next generation gravity missions

1. Introduction

Increasing global concentrations of greenhouse gases raise the ability of our planet to absorb solar energy and thus lead to globally rising temperatures. Higher atmospheric temperatures generally increase the capacity of the air to carry moisture, leading to potentially enhanced precipitation rates in several regions of the Earth. In addition to projections from coupled climate models more and more evidence is also emerging from satellite and in situ observations that changes in the terrestrial water cycle as triggered by modified precipitation pattern and intensities are already happening today [1]. Terrestrial water storage (TWS) is and possibly will be altering in terms of long-term linear wetting or drying trends [2], and increasing or decreasing seasonal amplitudes or time shifts in the seasonal cycle [3]. Furthermore, changes in the magnitude and occurrence frequency of extreme events [4] and interannual variations are expected [5]. Those changes pose a challenge for water management authorities engaged in balancing requirements on water consumption, renewable energy production, and flood control, which can only be met with a broad information basis provided by a well developed observing system.

Satellite observations of TWS are routinely available with global coverage from the Gravity Recovery and Climate Experiment (GRACE, in orbit from April 2002 to October 2017) [6], and its follow-on mission (GRACE-FO, in orbit since May 2018) [7,8]. The growing data record is increasingly being used for climate applications (see Tapley et al. [9] for a general overview) including for example the assessment of interannual variations in snow accumulation in Antarctica [10], and the relation of low-frequency variations in barystatic sea-level rise to intermittent increase in water storage at the continents [11]. Yet, the length of the time series is still limited and the attribution of changes to altering climate conditions is still difficult to conduct. In Rodell et al. [12] an attempt was made to classify regional TWS trends regarding their potential causes including climate change, and Jensen et al. [13] identified regions with coherent wetting or drying trends in climate model projections and the satellite record. Despite the progress achieved so far, in many regions overlaying long-term trends and interannual variations cannot be distinguished yet [13].

Next Generation Gravity Missions (NGGMs) are currently being prepared to extend the TWS data record with higher accuracy and spatio-temporal resolution. Different mission concepts are being considered that vary in terms of orbit design and resulting spatial-temporal ground track pattern [14]. NGGMs are supposed to improve the typical GRACE-type concept of two satellites (i.e., one satellite pair) following each other on the same orbit with an inter-satellite distance of about 200 km, thereby only observing the along-track component of the Earth's gravity field. A promising next-generation mission concept is a double-pair constellation being composed of two of such in-line pairs [15], a polar pair similar to the GRACE-type concept, and an independently operating second satellite pair with an inclination of 65–70 degrees. It was shown that this so-called Bender constellation results in a significantly improved error structure and accuracy compared to the classical GRACE-type concept [16,17].

Extensive end-to-end satellite simulations are typically performed to select a mission configuration and to demonstrate its potential value with respect to pre-defined user requirements [18]. Up to now, such simulation studies were mainly carried out for a few years to characterize the short-term performance of a mission [19], but hardly ever over more than a decade which would be important to assess the ability of missions to monitor climate variations. Simple error propagation from short-term simulations to long time periods does not provide adequate long-term performance estimates because the relative contribution of largely stochastic instrument errors and systematic errors (mainly resulting from temporal aliasing of short-period tidal and non-tidal signals) to the total error budget changes with increasing averaging period. Thus, for a realistic picture of the achievable performance and uncertainty characterization, long-term NGGM simulations are needed. Such simulations require realistic time series of the future evolution of TWS over several decades as input. Global coupled Earth System Models (ESMs) can provide information on the long-term development of TWS. Numerous different ESMs are taking part in the Climate Model Intercomparison Project Phase 6 (CMIP6) [20], and deliver projections of climate conditions until 2100 under the assumption of certain scenarios for the development of greenhouse gas concentrations. In this study, we therefore investigate land water storage related variables from CMIP6 models regarding their seasonal-to-interannual variability. Our goal is to characterize the most likely changes in TWS variability as seen by the multi-model ensemble in general, and to select a realistic model realization from the multi-model ensemble that can serve as input for long-term NGGM simulation studies. To achieve this, we

1. compare the variability of the TWS signal in GRACE and CMIP6 ESMs within the GRACE period (2002–2020) to demonstrate performance and identify shortcomings of the models;
2. analyze changes in the variability of TWS from model projections until the end of the century (2000–2100) and the consensus on such changes within the model ensemble;

3. perform a first step to assess the principle detectability of projected TWS changes with a GRACE-like gravity mission (with possibly higher sensitivity than GRACE, such as a potential double-pair NGGM); and
4. identify a representative model run from the ensemble of CMIP6 models which can serve as input data for NGGM simulations.

2. Materials and Methods

2.1. GRACE and GRACE-FO Data

To obtain a time series of global TWS grids from observations we make use of the ITSG-Grace2018 Level-2 data [21]. These data consist of 183 monthly solutions from the GRACE and GRACE-FO mission in the time period April 2002 to April 2020, which are given in the form of spherical harmonic coefficients of the gravitational potential up to degree and order 96. To account for the effect of geocenter motion the degree-1 harmonic coefficients provided by Sun et al. [22] based on Swenson et al. [23] are added. The c_{20} coefficient is replaced using a time series from Satellite Laser Ranging [24]. Furthermore, we consider glacial isostatic adjustment (GIA) by subtracting the ICE6G-D model [25] prior to our analysis. In order to reduce the anisotropic errors causing a striping pattern in the gravity solutions we apply a DDK3 filter [26]. Afterward, the spherical harmonic solutions are converted to equivalent water heights and evaluated on a global 2° × 2° geographical grid according to

$$TWS(\lambda,\theta) = \frac{M}{4\pi R^2 \rho_w} \sum_{n=1}^{n_{max}} \sum_{m=-n}^{n} \frac{(2n+1)}{(1+k'_n)} c_{nm} Y_{nm}(\lambda,\theta) \tag{1}$$

where λ and θ denote the spherical coordinates, M and R are the mass and the radius of the Earth, $\rho_w = 1000 \frac{kg}{m^3}$ is the density of water, k'_n denote the Load Love Numbers [27], c_{nm} are the filtered spherical harmonic coefficients of the gravitational potential, and $Y_{nm}(\lambda,\theta)$ are the surface spherical harmonic functions. The result is considered to represent water storage changes on land. This assumption is not entirely true everywhere, because residual tectonic signals from GIA [28], post-seismic deformation after large earthquakes [29,30], or residual atmospheric mass variability [31] may overlay the TWS signal in certain regions. In the time series of GRACE TWS grids, individual months are missing due to repeat-orbit constellations or instrument outages especially towards the end of the GRACE mission. These individual missing months (21 in total) were linearly interpolated to obtain a continuous time series for subsequent signal decomposition (Section 2.3). The data gap of 11 months between the end of the GRACE mission and the start of the GRACE-FO mission was not interpolated but excluded from the study.

Together with the gridded TWS values, we derive their uncertainties. To obtain realistic error estimates, the full error-covariance matrices of the spherical harmonic coefficients are used. Here, we make use of an exemplary GRACE error-covariance matrix of a particular month (2008/01) to propagate the uncertainty of the potential coefficients including their correlations to the gridded TWS values. The resulting uncertainty grid is considered to be representative for the GRACE observational accuracy and kept constant over time when deriving accuracies for individual signal components (Section 2.3). This assumption is justified by the fact that GRACE errors do not scale with the signal, but are mainly driven by sensor noise (i.e., the accelerometer and the inter-satellite ranging system) as well as background model errors [19]. The main variations not being considered with this assumption are the impacts of the changing satellite ground track due to the drifting GRACE orbit and the instrument degradation towards the end of the GRACE mission. The former can be considered a minor issue, except for a few specific months affected by a deep orbit resonance, i.e., a very short repeat period and thus a significantly degraded spatial resolution, and the latter do not represent typical errors of a GRACE-like mission as we intend to assess here.

2.2. CMIP6 Model Data

The ESMs in CMIP6 provide total soil moisture content (mrso) and surface snow amount (snw) as land water storage related variables. In contrast to the GRACE observations, which capture all parts of TWS from the deepest aquifer up to the surface as an integral signal, CMIP6 models do not include explicit groundwater modeling, surface water representation, or mass changes from ice sheets, glaciers, and ice caps. Furthermore, human interventions such as groundwater abstraction, irrigation or dam building are not considered. Although not explicitly modeled, parts of the groundwater variability may be contained in the mrso variable, because the mass transport to ocean and atmosphere is limited and the water balance is approximately closed by most of the models [32]. However, as the interaction processes between groundwater, soil, and surface water are not considered, there might be systematic errors in the representation of modeled TWS variations, which can only be reduced by further model development or accounted for by improved methods for separating groundwater from the remaining TWS signal in the observations. The depth of the soil moisture layers in ESMs varies depending on the model from just a few to several tens of meters, and is spatially invariant. Therefore, in several regions the ESMs do not represent the entire TWS variability as seen by GRACE. Nevertheless, ESMs provide a valuable proxy for the expected evolution of TWS variability. The impact of differences between observed and modeled TWS on different signal components is extensively assessed and discussed in this study (Section 3.1). Furthermore, we identified regions where discrepancies between TWS and mTWS caused by surface water storage changes, groundwater abstraction, or glacier mass changes might influence the results (see Supplementary Material, Section S1). In these regions (about 11% of the land area) the results of the comparison between models and observations have to be interpreted with care.

The CMIP6 data base is still growing as more modeling groups are providing their results. At the time of writing 25 models provide monthly output of global mrso and snw grids for the historical experiments available for the time span 1850–2011 and the Shared Socioeconomic Pathway 5–8.5 (SSP585) projections that are based on scenarios of the evolution of greenhouse gas emissions and cover the time span 2012–2100. For most of the models several (up to 50) simulation runs are available that were produced by slightly varying the initial conditions. Each model run can be considered to be a possible projection of future climate conditions. Together, all model runs build an ensemble, and in case of several models, a multi-model ensemble is analyzed. Each ensemble member data is processed as follows: We calculate the sum of mrso and snw and refer to it as modeled TWS (mTWS). We concatenate the monthly mTWS grids of the historical and the SSP585 experiments to cover the time period of the observations and the future evolution until the end of the century. Afterward, the mTWS grids are remapped to a common 2° global resolution.

The 25 models that currently provide mrso and snw data are not all fully independent from each other, but instead are partly improvements or extensions of each other, or share central elements, such as land, atmosphere, or ocean sub-models. In order to obtain unbiased results when analyzing multi-model averages, we reduced the ensemble by omitting all highly correlated experiments as outlined in the Supplementary Material (Section S2). In the end, 17 models with altogether 105 ensemble members remain for the analysis in this study. Detailed information and references for the 17 models remaining (Figure S2) can be accessed, e.g., via https://esgf-data.dkrz.de/projects/cmip6-dkrz/.

A standard procedure to comprehend information from model results is the calculation of a multi-model average. Here, for robustness, we choose the median instead of the arithmetic mean. Thus, the median grid for each time step is obtained by calculating for each grid cell the median of all $N = 105$ model values. In order to give each model the same weight, regardless of the number of ensemble members belonging to it, we compute the *weighted* multi-model median. For clarity, here we denote it with *unscaled* weighted multi-model median (unscaled MMMed) to distinguish it from the *scaled* MMMed introduced in Section 2.5. The weights $w_i = 1/K$ assigned to each model run are calculated as the reciprocal value of the number K of ensemble members per model, with K varying between 1 and 50. For example, if a model has three members, each of them gets a weight of 1/3. As a result, all weights w_i sum up to the number of models $V = \sum_{i=1}^{N} w_i$ (here 17). The weighted median is defined as the element $\overline{m} = x_k$ from N ordered elements $x_1 \ldots x_N$ with corresponding weights $w_1 \ldots w_N$ where

$$\sum_{i=1}^{k-1} w_i \leq V/2 \text{ and } \sum_{i=k+1}^{N} w_i \leq V/2. \tag{2}$$

This means that the element at the index k where the cumulative sum of the weights is (for the first time) larger than 50% of the total sum of the weights V is selected. As a measure of uncertainty for the unscaled MMMed, we compute the model spread as the weighted standard deviation of the ensemble members:

$$\sigma_x = \sqrt{\frac{1}{V} \sum_{i=1}^{N} w_i (x_i - \overline{m})^2} \tag{3}$$

2.3. Signal Decomposition

To analyze different constituents of the TWS signal, we decompose it into a long-term component, a seasonal cycle (annual and semiannual), and sub-seasonal variations. The long-term component is further separated into a linear trend and interannual variations [33,34].

$$\text{TWS}_{\text{total}} = \underbrace{\text{TWS}_{\text{long-term}}}_{\text{TWS}_{\text{linear}} + \text{TWS}_{\text{inter}}} + \text{TWS}_{\text{seas}} + \text{TWS}_{\text{sub}} \tag{4}$$

The linear trend and the seasonal cycle is estimated from the full time series in terms of a least-squares adjustment. In order to identify potential changes in the annual cycle over the next decades we co-estimate linear trends for the amplitude and the phase instead of keeping them constant over time. The following model is fitted to the TWS time series:

$$\hat{\text{TWS}}(t) = a + b \cdot t + (c + c' \cdot t) \cdot cos(\omega t - (d + d' \cdot t)) + e \cdot cos(2(\omega t - f)) \tag{5}$$

with parameters for bias (a), linear trend (b), annual and semi-annual cycle (c,c',d,d',e,f). After removing the linear trend and the seasonal cycle from the total signal, the sum of interannual and sub-seasonal variations remains. To distinguish the two components we apply a Butterworth filter with 12 months filter length. The decomposition of the total signal is displayed in Figure 1 for an exemplary mTWS time series. The Supplementary Material (Section S3) contains for the same position the corresponding GRACE TWS time series and its decomposition.

The decomposition is performed for each land grid cell, for each of the 105 CMIP6 ensemble members, for the unscaled MMMed model time series, and for the GRACE data set. Furthermore, it is applied on two different time spans, 2002/04–2020/04 (the GRACE and GRACE-FO period) and 2000/01–2100/01 (only for the model data). To ensure consistency for the comparison between GRACE and models, the 11 month data gap between GRACE and GRACE-FO was also excluded from the model time series prior to decomposition. As information on the accuracy of the grid values for the GRACE and the unscaled MMMed model time series is available, we can strictly propagate these during the parameter estimation in Equation (5), ending up with standard deviations σ_a, ..., σ_f for the different signal components for each grid cell. For the 105 individual model runs we have no information on their accuracy, hence, no error estimates for their signal components, so that only the ensemble spread is used to characterize model uncertainty.

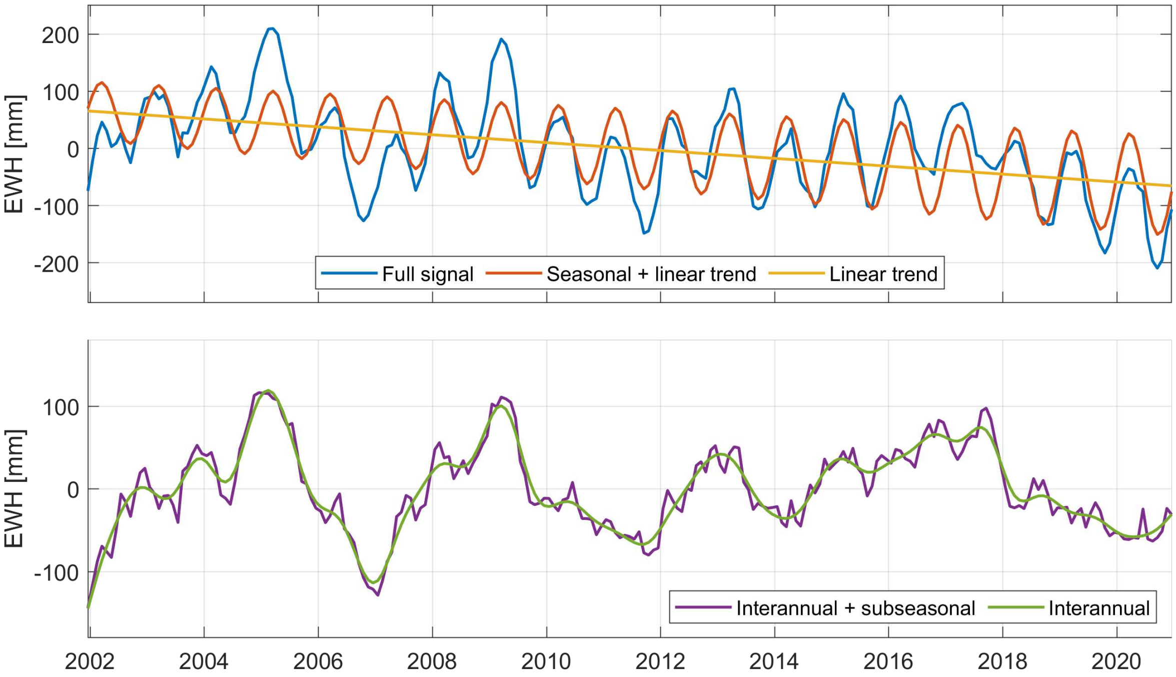


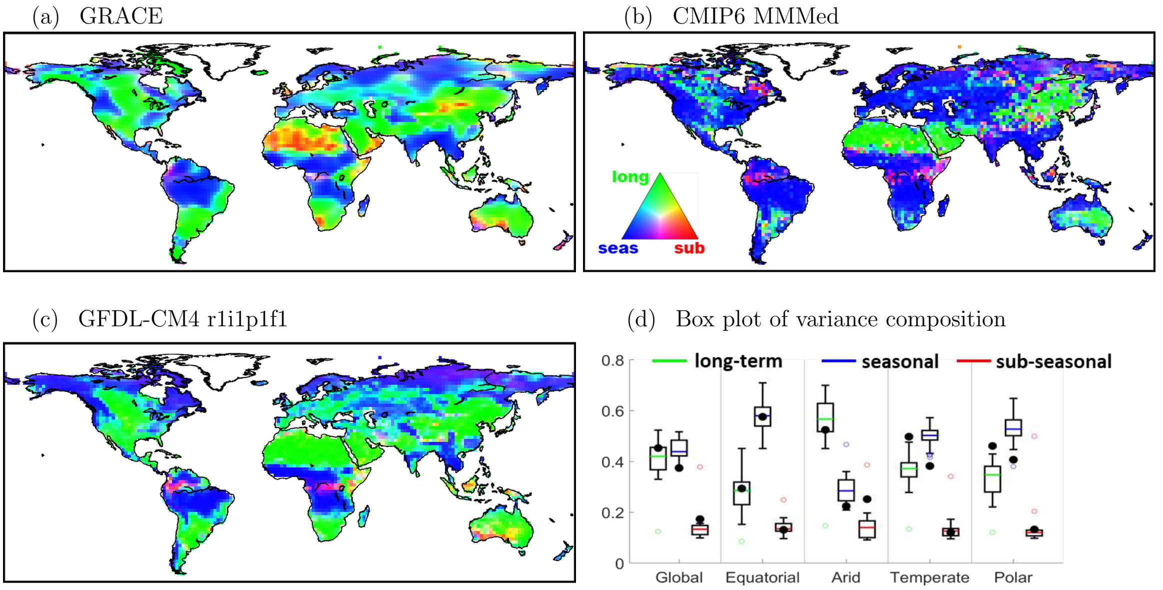
Figure 1. Example for the decomposition of a mTWS time series into linear trend, seasonal, subseasonal and interannual signal. The location is 13 °E and 52.5 °N (Potsdam, Germany), the model is GFDL-CM4 (r1i1p1f1 run).

2.4. Relative Importance of Signal Components

We now split TWS and mTWS signals into different temporal components as described above. For each signal part in each grid cell, we compute the signal variance over the GRACE time span and relate it to the variance of the total signal. The fractional variance composition for the long-term, seasonal and sub-seasonal signal is displayed in Figure 2. The colors in Figure 2 are assigned by mixing the red, green, and blue (RGB) color values according to the fractions of the variance components, meaning that pure blue would indicate a perfect seasonal signal with no long-term or sub-seasonal components. Pure green characterizes no seasonal or sub-seasonal variations, and pure red a location with only sub-seasonal variability. A signal with equal long-term, seasonal and sub-seasonal variance would be displayed white, and other mixtures with the colors in-between. The variance component analysis of the GRACE TWS time series (Figure 2a) reveals that many regions at moderate latitudes are particularly affected by long-term variations, so that water availability as represented by TWS is particularly modified from long-term natural (or anthropogenic) climate variability.

Repeating the same variance component analysis for the unscaled MMMed mTWS time series of the 17 models instead results in a pattern largely dominated by the seasonal variance (Figure 2b), which is very distinct from the results obtained from GRACE. This discrepancy reveals a caveat of using the unscaled MMMed for comparison with observations: Climate models are able to represent interannual and sub-seasonal TWS variations in a statistical manner only. This implies that the exact timing of the occurrence of troughs and peaks in the time series is random so that model runs and observations are not directly comparable on time series level in terms of, e.g., correlation or RMSD. Hence, time series of interannual and sub-seasonal TWS variations from different model runs can match only regarding their magnitude and frequency, but not at specific points in time. Therefore, when building the model average, the interannual and sub-seasonal variabilities that are contained in the individual model runs are not maintained but largely smoothed out. As a consequence, mainly seasonal signals remain in the MMMed. Thus, it is not feasible to directly compare the signal variability of the unscaled MMMed time series with the observations, but we have to revert to investigating the variability of individual ensemble members.

Exemplarily, in Figure 2c the variance composition of one specific run (r1i1p1f1) of the GFDL-CM4 is shown, which exhibits much more similarities with the observational pattern in Figure 2a than the unscaled MMMed pattern. Hence, for the remaining sections of the paper, we first perform the signal decomposition for all 105 ensemble members and afterward calculate the MMMed and its respective standard deviation for the individual components (e.g., annual amplitude) according to Equations (2) and (3).


(a) GRACE

(b) CMIP6 MMMed

(c) GFDL-CM4 r1i1p1f1

(d) Box plot of variance composition

Figure 2. Distribution of the total variance into long-term, seasonal and sub-seasonal variability for the time span 2002/04–2020/04 for (**a**) the GRACE TWS observations, (**b**) the MMMed of the simulated mTWS, (**c**) an exemplary individual model run (GFDL-CM4 r1i1p1f1). (**d**) Box plot of the variance composition within the model ensemble for the global average and averaged over different climate zones. Black dots denote the respective result for the GRACE observations.

To achieve a comprehensive overview over the variance composition in the whole model ensemble, we aggregate the fractional variance components as follows: For each of the 105 ensemble members we compute for each land grid cell the variance composition (as before: three fractional values for long-term, seasonal, and sub-seasonal, adding up to 1.0). Afterward, we compute for each ensemble member the global land average of the individual component fractions, resulting in 105 times 3 values representing the model range of variance compositions for the global mean. This range of variance compositions is displayed in the left of Figure 2d. The boxes represent the median (colored line), the 25% and 75% percentile (bottom and top edge of box) and the most extreme data points not considered outliers (whiskers) in the data set of ensemble members. The outliers are displayed separately with colored circles. For comparison we also compute the global land average for the three variance components of the GRACE record (black dots in Figure 2d). Furthermore, we repeat the averaging of the component fractions for different Köppen-Geiger climate zones [35] (further diagrams in Figure 2d). Please note that in all global (and regional) land averages provided in this study, Greenland and Antarctica are excluded from the computations as mass change in these regions is dominated by ice mass variations, which are not represented by the ESMs. Comparing the model results to the observations (Figure 2d) we note that for the fractional long-term component the models are close to GRACE in equatorial and arid regions but underestimate it in temperate and polar regions. The fraction of the annual cycle, however, is mostly overestimated by the models except for equatorial regions. The fit of the sub-seasonal variance fraction is good for equatorial, temperate and polar regions. In arid regions the sub-seasonal variance fraction in the models is smaller than in the GRACE observations. One reason for this is probably inherent noise in the gravity data causing a low signal-to-noise ratio in arid regions with low signal variability. In equatorial regions the variance composition into long-term, seasonal and sub-seasonal signal fits remarkably well to the observations, whereas it is more discordant in the other regions.

2.5. Building and Rescaling the MMMed for Measures of Variability

In the previous section we pointed out that the computation of an unscaled MMMed grid time series from all 105 ensemble members results in a reduced interannual and sub-seasonal variability. Thus, to analyze the variability in the model ensemble, we separately decompose all 105 ensemble member time series and define the amplitude and phase of the annual cycle as well as the Root Mean Square (RMS) of the interannual signal component as the measures of signal variability to be investigated. Only afterward we then compute the MMMed grid over these measures. By this, we maintain the overall variability of the model ensemble and obtain for each signal variability measure a global pattern reflecting the best estimate for its spatial distribution.

However, there is a second issue arising from building an unscaled MMMed on grid cell level: as the global patterns differ for different model runs, the median smooths out extreme values in each grid cell. Thus, the range of values in the unscaled MMMed grid is much smaller than the actual range of values in the individual ensemble members (and the observations). For illustration, Figure 3 displays the empirical cumulative density functions (ECDFs) for the 5–95% percentiles calculated from the land grid cells of all model runs for one of the measures we investigate (interannual RMS change, Section 3.2.2). Additionally, the ECDF of the unscaled MMMed is shown (thick red line), which has a much smaller range and thus cannot be regarded as a realistic representative for the variability. Thus, we adjust the range of values in the MMMed as follows: we compute the weighted multi-model median of all 105 ECDFs (thick black line in Figure 3) to obtain a best estimate for the range of simulated values, and divide it for each percentile by the ECDF value of the unscaled MMMed (thick red line). Each grid cell value of the unscaled MMMed map is then multiplied with the factor belonging to its respective percentile. As a result, the ECDF of the *scaled* MMMed equals the MMMed of the 105 ECDFs of the ensemble members (thick black line). Consequently, also the scaled MMMed standard deviation is adjusted by this factor. The scaling of the MMMed ensures that not only the global pattern but also the range of values states a representative estimate of the model results. Thus, in the rest of the paper, all comparisons (as far as referring to a model average) are performed with respect to the *scaled* MMMed. If only MMMed is written, it always means the scaled MMMed.

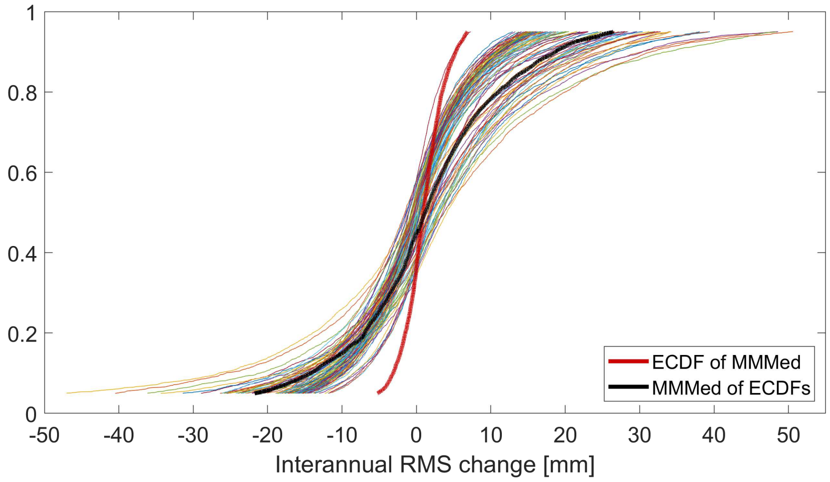


Figure 3. Example for different range of values for the ensemble members and the MMMed.

3. Results

In Section 3.1 we compare the CMIP6 mTWS model results to GRACE TWS observations for the time span 2002/04–2020/04 in terms of annual cycle (amplitude and phase) and interannual variations as derived from the signal decomposition. From the models the (scaled) MMMed as a best estimate for the respective signal component is used in the comparison. Afterward, in Section 3.2 the future development of the annual cycle and interannual variations for the time span 2000/01–2100/12 is assessed by the MMMed (serving as expectation value) of the respective signal component changes. Furthermore, the MMMed long-term linear trend is evaluated. We also discuss the consensus of the models on the projected changes. In Section 3.3, in order to provide a first hint about the detectability of projected TWS annual cycle changes, the MMMed annual cycle changes over 30 years are compared to the accuracy of GRACE and possible NGGM observations. In the last part (Section 3.4) we select a specific model run from the ensemble of 105 members that can serve as input for NGGM simulation studies. In this selection process, we consider both the fit of the model run to GRACE observations in the GRACE time span (for annual cycle and interannual variations) and to the MMMed in the projected time span (for changes in amplitude, phase, interannual variations, and linear trends). An overview of the different evaluation time spans and measures discussed in this section is given in Table 1.

Table 1. Overview over the different signal components analyzed for different time spans.

2002/04–2020/04 **Comparison of MMMed to GRACE**	**2000/01–2100/12** **MMMed Values & Consensus**
annual amplitude (Section 3.1.1)	annual amplitude change (Section 3.2.1)
annual phase (Section 3.1.1)	annual phase change (Section 3.2.1)
RMS of interannual signal (Section 3.1.2)	change of interannual RMS (Section 3.2.2)
	linear trend (Section 3.2.3)

3.1. Current TWS Variability in CMIP6 Models and GRACE

3.1.1. Seasonal Cycle

Here we compare GRACE and CMIP6 models first regarding the annual amplitude. The global patterns (parameter c of Equation (5)) from GRACE and the (scaled) MMMed annual amplitudes from 17 CMIP6 models (Figure 4a,b) are visually similar and feature a pattern correlation of 74%. The pattern correlation of two grids is calculated by arranging (for each grid separately) the values of all land grid cells into a vector and computing the Pearson product-moment correlation coefficient of the two vectors. The correlation coefficient can be regarded as a measure of pattern similarity for the two grids.

To compare the model results and GRACE, we compute the ratio between the CMIP6 MMMed amplitudes and the GRACE amplitudes (Figure 4c). The more the ratio differs from 1 in a certain grid cell, the less the amplitudes from models and GRACE correspond in this grid cell. To objectively rate the magnitude of the deviation, the uncertainty of the ratio has to be considered. The standard deviation of the ratio is derived from the model spread of the MMMed amplitude (Equation (3)) and the standard deviation of the GRACE amplitude via variance propagation. We find that for 66% of the global land area (without Greenland and Antarctica) the amplitude ratio deviations from 1 lie within the standard deviation of the ratio. Large parts of the deviations that exceed the error bounds (stippled regions in Figure 4c) are located in arid regions of Northern Africa and the Arabian Peninsula, which generally exhibit a small annual cycle of TWS, and thus the relative uncertainties are larger. To facilitate a more detailed analysis of the CMIP6 amplitudes and their relation to GRACE we calculate the CMIP6 signal-to-noise ratio (SNR) for the annual amplitude (MMMed amplitude divided by the model spread of the MMMed amplitude, Figure 4d). Apart from regions with very small values (discussed below) the SNR is generally higher in the northern hemisphere than in the southern hemisphere. We attribute this to the fact that for many observational records (especially in situ observations) the coverage in the northern hemisphere is denser and thus the calibration of existing climate models is askew respectively, leading to a reduced model spread for northern regions.

In the arid regions of Northern Africa, the Arabian Peninsula, and in large parts of China and Mongolia, the SNR is very small or even below 1. In these regions, the annual cycle is generally not very distinctive (Figure 4b), as can also be seen from the composition map in Figure 2c: the dominating component of the time series in the regions of small SNR is the long-term or sub-seasonal signal. Thus, meaningful results for the annual cycle cannot be expected in these regions. We decided to exclude regions with a SNR of the CMIP6 amplitudes below 1 (24% of the global land area) prior to a regional analysis of the amplitude deviations between CMIP6 and GRACE, because otherwise the statistical measures would be largely distorted.

While putting regions with SNR below 1 aside, the median overestimation of the amplitude with respect to GRACE (area-weighted median of all ratios in Figure 4c above 1) is 1.38 and the median underestimation (area-weighted median of all ratios below 1) is 0.79. The land area where the annual amplitudes are overestimated by the models is larger than the area where it is underestimated (58% vs. 42%). This proportion varies depending on the climate zone. For the regional analysis we access the Köppen-Geiger climate classification [35] and compute the median over- and underestimation for four different regions characterized by arid, equatorial, temperate, and polar climates. All numbers of the regional analysis are given in the Supplementary Material (Tables S1–S7). The regions with SNR below 1 are mainly arid: here, 50% of the area is excluded, whereas in the other climate regions this fraction is only 10%. The proportion of overestimation to underestimation in equatorial regions is 51% to 49%, i.e., a more even proportion compared to the global values. In contrast, the annual amplitude is mainly overestimated (65%) in polar regions. A tendency of underestimation in tropical regions and overestimation in northern regions was also found by Scanlon et al. [34] for a selection of hydrological models and land surface models, especially for the CLM-5.0 model that is the land surface component in two ESMs investigated in this study (CESM2-WACCM, NorESM2-LM). Particularly large amplitude underestimation of CMIP6 amplitudes compared to GRACE that exceeds the model uncertainty occur in the Amazon and the Ganges-Brahmaputra basin. We suppose that the water holding capacity is bound-limited in models, leading to an overly strong runoff of excess water. In addition, soil moisture memory is often too short in models especially in equatorial regions, preventing the accumulation of water to its real storage extent. The overestimation of the amplitude by the models in the north might be related to overestimated snow storage in winter and evapotranspiration in summer, thereby simulating an overall increased annual amplitude [34].

In addition to the amplitude we also investigate the MMMed phase of the annual cycle (i.e., the month of the annual TWS maximum, derived from parameter d of Equation (5)) and compare it to the GRACE-derived phase (Figure 5a,b). Please note that phase values are cyclic and cannot be assumed to be normally distributed. A scheme for the calculation of a weighted median from phases and its standard deviation is described in the Supplementary Material (Section S5). As the computation of a ratio of phases between models and GRACE is not feasible, we only calculate the differences between the MMMed phases and the GRACE phases (Figure 5c) to compare the two grids. The accuracy of the phase differences is computed by error propagating the model spread of the MMMed phase and the standard deviation of the GRACE phase. In 74% of the global land area the differences between the GRACE and the model phases do not exceed the standard deviation of the difference (Figure 5c).

Particularly large differences occur in the arid regions of Northern Africa and the Arabian Peninsula, as well as in those parts of China and Mongolia that exhibit a very small SNR of the amplitude (Figure 4d), hence a weak annual cycle that prevents meaningful phase estimates. Therefore, we again exclude regions with an amplitude SNR below 1 from the further analysis. The remaining differences are mainly positive (72% of land area), meaning that the observed annual cycle is slightly lagged behind the modeled cycle. The median positive phase shift (models earlier, 72%) is 0.50 months, and the median negative phase shift (models later, 28%) is -0.32 months. The mainly positive phase shift between observations and models might be related to missing groundwater processes in CMIP6 models causing an underestimation of the water residence time in the soil, hence less time for storage accumulation and consequently an earlier saturation of the maximum storage. The tendency of the models to precede the observed annual cycle was also found for the hydrological and land surface models investigated by Scanlon et al. [34]. It is larger in the polar climate zone (82%) than in the equatorial zone (62%).

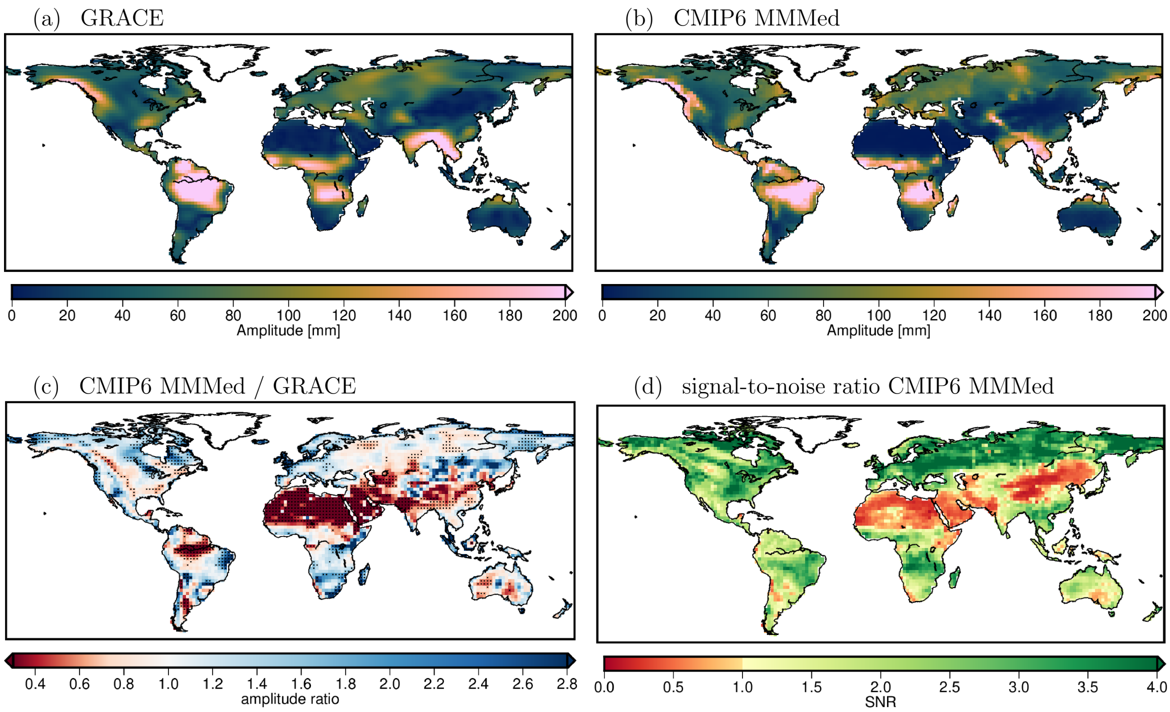


Figure 4. (**a**) GRACE TWS annual amplitude and (**b**) scaled MMMed mTWS annual amplitude for the time span 2002/04–2020/04. (**c**) Ratio of (**b**) and (**a**). Stippling indicates regions where the deviation from 1 exceeds the standard deviation of the ratio. (**d**) MMMed mTWS annual amplitude signal-to-noise ratio.

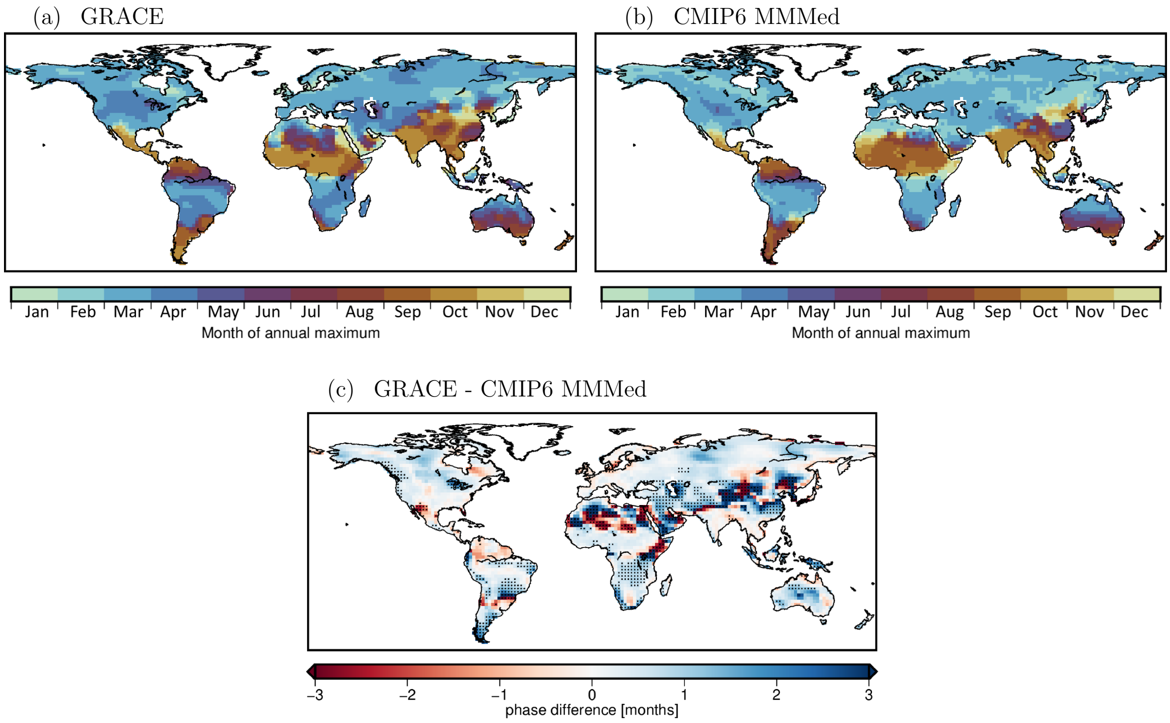

(a) GRACE
(b) CMIP6 MMMed
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month of annual maximum
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Month of annual maximum
(c) GRACE - CMIP6 MMMed
−3 −2 −1 0 1 2 3
phase difference [months]

Figure 5. (**a**) GRACE TWS month of the maximum of the annual cycle and (**b**) MMMed mTWS month of the maximum of the annual cycle for the time span 2002/04–2020/04. (**c**) Difference of (**a**,**b**). Stippling indicates regions where the difference exceeds the standard deviation of the difference.

3.1.2. Interannual Anomalies

In several parts of the world long-term variations (linear trends plus interannual variations) in TWS are the dominating part of the signal variability (Figure 2). The direct comparison of linear TWS trends from GRACE and ESMs is difficult because the models do not include all physical processes (e.g., surface water, glacier, and anthropogenic groundwater change) that contribute to observed TWS, and therefore are not able to fully represent long-term trends everywhere. Furthermore, over a time span of less than 20 years trends are largely influenced by interannual variability [13], so that we focus on interannual variations only. As elaborated in Section 2.3, model time series of interannual variations are not directly comparable due to the random occurence of such variations in the models. Therefore, we chose the RMS of the interannual variability over 2002–2020 as a measure for comparison. Similarity of the RMS between models and GRACE over a 18-year time span would prove that the models are able to capture the general range of observed interannual variability independent of the exact evolution of the time series. Indeed, the global patterns of the GRACE interannual RMS and the MMMed interannual RMS (Figure 6a,b) feature similarities in many regions of the world. The global pattern correlation for the MMMed with GRACE is 64%, which is lower than the correlation for the annual amplitude, but nearly as high as the correlation of the MMMed with its individual model ensemble members (65 ± 15%), which highlights the large intermodel spread of the interannual TWS variability in ESMs.

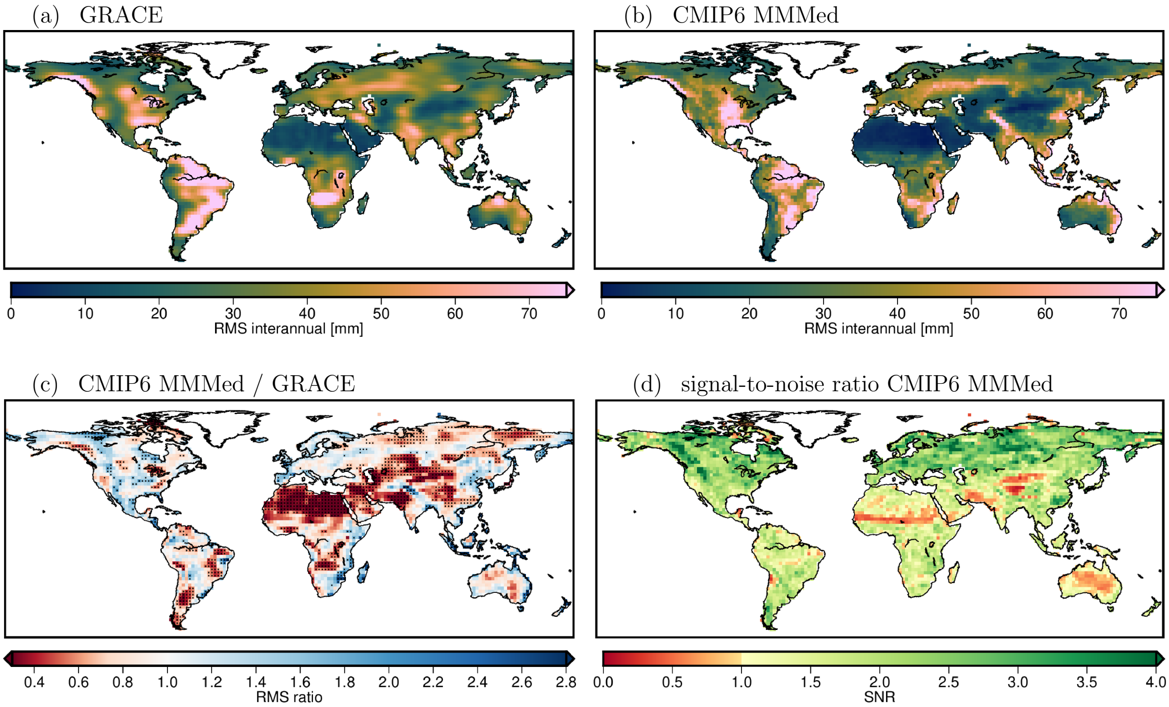


Figure 6. (**a**) GRACE TWS interannual RMS and (**b**) MMMed mTWS interannual RMS for the time span 2002/04–2020/04. (**c**) Ratio of (**b**) and (**a**). Stippling indicates regions where the deviation from 1 exceeds the standard deviation of the ratio. (**d**) MMMed mTWS interannual RMS signal-to-noise ratio.

As we did when analyzing the annual amplitude, we also compute the ratio of the MMMed interannual RMS and the GRACE interannual RMS (Figure 6c). As before, the accuracy of the ratio is derived by variance propagation of the respective MMMed and GRACE accuracies. Regions, where the deviation of the ratio from 1 exceeds its standard deviation, correspond to 31% of the land area (stippling in Figure 6c), mainly in South America, Central and South East Asia, and in large parts of Africa. The interannual RMS is underestimated compared to GRACE observations in 60% (regions of SNR < 1 from Figure 6d excluded) of the global land area. This proportion is relatively constant for the different climate zones (54% to 68%). There are also regions where the interannual RMS in models is overestimated compared to GRACE (40%), mainly in coastal areas (North and West Europe, Australia, East Coast of Africa, North Coast of Canada) and in the mountainous areas in North America. It would be interesting to explore reasons for those discrepancies in cooperation with climate scientists that are developing the CMIP6 models. The model signal-to-noise ratio, i.e., the MMMed interannual RMS divided by its standard deviation (Figure 6d), is generally smaller than for the annual amplitude, which confirms that the representation of the seasonal cycle in ESMs is more reliable than the representation of interannual variations. This is an expected result as the underlying physical processes driving interannual fluctuations are more complex and less known than the annual cycle, causing its modeling to be more challenging.

3.2. Projected Change in TWS Variability until 2100

Model projections until 2100 provide valuable information about the future development of TWS. We therefore analyze changes in the annual cycle and interannual variations of the CMIP6 mTWS signal for the time span 2000/01–2100/12. Additionally, we analyze the long-term linear mTWS trend. The analysis is done for the (scaled) MMMed taken as expectation value for the future development of the climate.

3.2.1. Seasonal Cycle Changes

Centennial changes in the annual mTWS amplitude are obtained by accessing parameter c' in Equation (5) from the model signal decomposition over 2000/01–2100/12. The (scaled) MMMed of the amplitude changes exhibits values on the order of several mm EWH per decade (from -2.75 mm/yr to 1.83 mm/yr, with a median absolute value of 0.11 mm/yr), which are substantially varying locally (Figure 7).

In 45% of the land surface more than three quarters of the models (13 or more of 17, $\geq$76%) agree on the direction of the amplitude change. These regions are stippled in Figure 7 and are referred to as high consensus regions in the following. As in high consensus regions many models are concordant about the direction of the change we consider the results in these regions as particularly reliable. In many regions a high model consensus corresponds to a large magnitude of the amplitude change (e.g., in Europe and Northeast Asia, Canada) and analogously low consensus goes along with small changes. However, there are exceptions from this rule. For example, there is only low model consensus about changes in amplitudes in large parts of South America although the mean amplitude changes especially in the Amazon basin are among the largest signals. Also in Central Africa the decreasing amplitude is not supported by high consensus. Conversely, there are also regions of small amplitude changes but high model consensus. These often also feature a small seasonal cycle, e.g., in Northern Africa and in Central Asia.

According to the models, in the majority of the land area (56%) the seasonal amplitude will increase until 2100, with a median of 0.12 mm/yr. The median of the amplitude decrease (44% of the land area) is -0.11 mm/yr. The area proportion of positive to negative changes is even more pronounced when restricting to high consensus regions (66% increasing amplitude with median 0.21 mm/yr, 34% decreasing amplitude with median -0.26 mm/yr). Furthermore, the distribution of amplitude changes depends on the climate zone, e.g., the proportion of increasing vs. decreasing amplitudes is more balanced in the equatorial zone (51% vs. 49%), and more distinct in the polar zone (69% vs. 31%) compared to the global distribution. The relatively strong amplitude increase in polar regions (median of 0.18 mm/yr) originates mainly from the soil moisture component of the ESMs, as can be seen from Figure 7b,c that show the amplitude changes for the mrso and snw variable separately. In the snow component a decrease in the annual amplitude is projected with high consensus, which is related to generally rising temperatures and reduced snow accumulation until 2100. The increase of the soil moisture amplitude might be due to increased evapotranspiration in a warmer climate therefore reducing water storage during summer.

Together with the linear trend of the annual amplitude we also estimate for each grid cell (and for each ensemble member) a linear trend of the phase of the annual cycle (parameter d' in Equation (5)). These linear trends can be converted to a MMMed total phase shift from 2000 to 2100 (Figure 8).

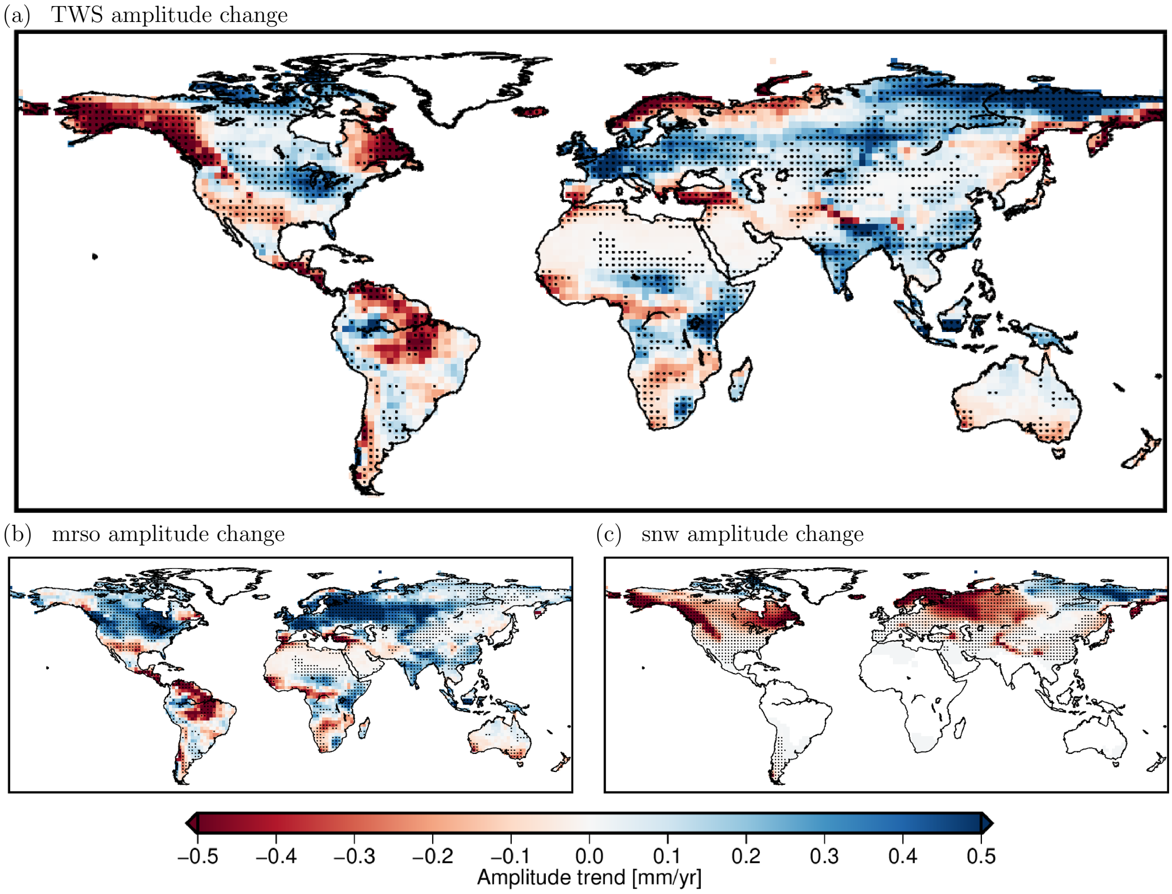


Figure 7. (**a**) MMMed mTWS annual amplitude change over 2000–2100. Stippling indicates regions where 13 or more of 17 models (i.e., ≥76%) agree on the sign of the amplitude change. (**b**,**c**) same as (**a**) but for mrso and snw.

In 37% of the land surface more than three quarters of the models (13 or more out of 17) agree on the direction of the phase shift. For the majority of the land area (55%) the models project a positive phase shift, i.e., the maximum of the annual cycle is reached later in 2100 than in 2000 (median of 0.39 months, approx. 12 days). The median of the 45% land area where the maximum is reached earlier is −0.35 months (approx. 11 days). The tendency to a later annual cycle is particularly strong in the equatorial zone (75% later, with a median of 0.49 months). Especially in Africa large parts of the continent will experience a substantial later peak of the annual TWS maximum according to a majority of the models. This is related to a later onset of the rainy season which was identified already in CMIP5 models by Dunning et al. [36] who attribute this to a position shift in the tropical rain belt and increasing strength of the Saharan heat low.

In the polar climate zone the area proportion of positive and negative phase shifts is opposite to the global (only 41% later and 59% earlier). We relate this to a generally shorter accumulation period of snow and an earlier onset of thawing due to higher air temperatures. When splitting the mTWS signal into the soil moisture and the snow component (Figure 8b,c) it is striking that in large parts of the polar climate zone both the mrso and the snw MMMed exhibit a strong negative phase shift (earlier reach of maximum in 2100), whereas the sum of both (mTWS MMMed) has a small positive phase shift. This is a result of the interference between the mrso and snw signals which mostly show growing mrso amplitudes and shrinking snw amplitudes in polar regions (compare Figure 7b,c) and are shifted in their respective phases. An example for this effect is given in the Supplementary Material (Section S6). This finding highlights the importance of an accurate modeling of all individual TWS components and their relative magnitude and phase, thereby underscoring once more the importance of satellite-based surface mass observations for guiding the numerical modeling of the global water cycle.

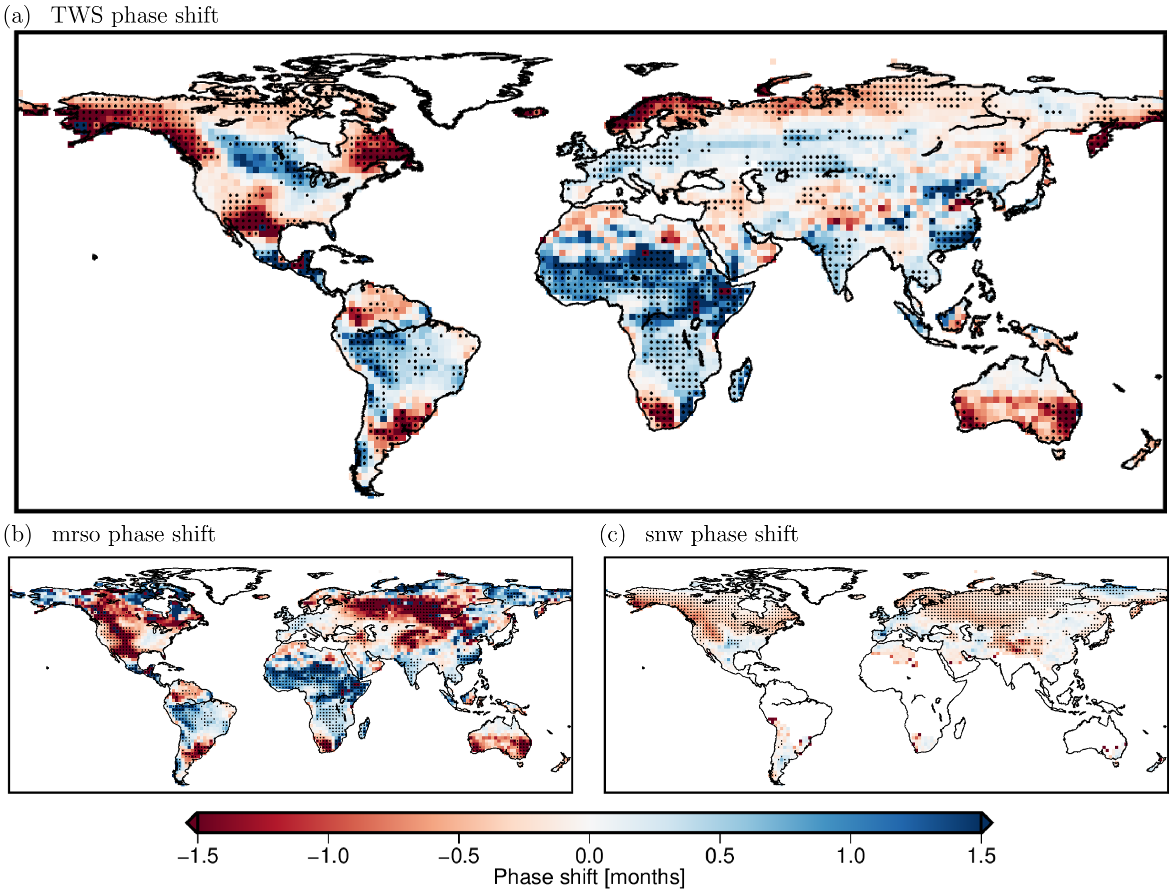


Figure 8. (**a**) MMMed mTWS phase shift of the annual cycle over 2000–2100. Stippling indicates regions where 13 or more of 17 models (i.e., ≥76%) agree on the sign of the amplitude change. (**b**,**c**) same as (**a**) but for mrso and snw.

3.2.2. Interannual Anomaly Changes

Changes of the RMS of the interannual signal over 2000–2100 were identified by computing the MMMed of the differences between the interannual RMS over two time spans, one at the end (2082–2100) and one at the beginning (2002–2020) of the investigated time period (Figure 9).

A slight increase of interannual variability is projected by the models for the majority of the land area (54%) with a median value of 7.02 mm. The median negative RMS difference is −5.75 mm (in 45% of the land area). However, the general pattern of the MMMed interannual RMS change is not as distinct as it is for changes in the annual cycle (Section 3.2.1), and in only 23% of the land surface a high model consensus (≥76%) is found. This is probably related to the larger intermodel spread for interannual variations compared to the seasonal signal (cf. Section 3.1.2). The proportion of land area with positive vs. negative RMS changes is quite similar for all climate zones (between 50% and 57% positive), thus no clear regional dependence of the development for interannual variations can be identified. However, when splitting the mTWS signal into its components soil moisture and snow (Figure 9b,c) a clear decrease of the snow interannual variability (supported by more than three quarters of the models) is found for all snow covered regions (except for a small patch in North East

Siberia), whereas the polar zone in the soil moisture projection is largely dominated by an increase of interannual variability.

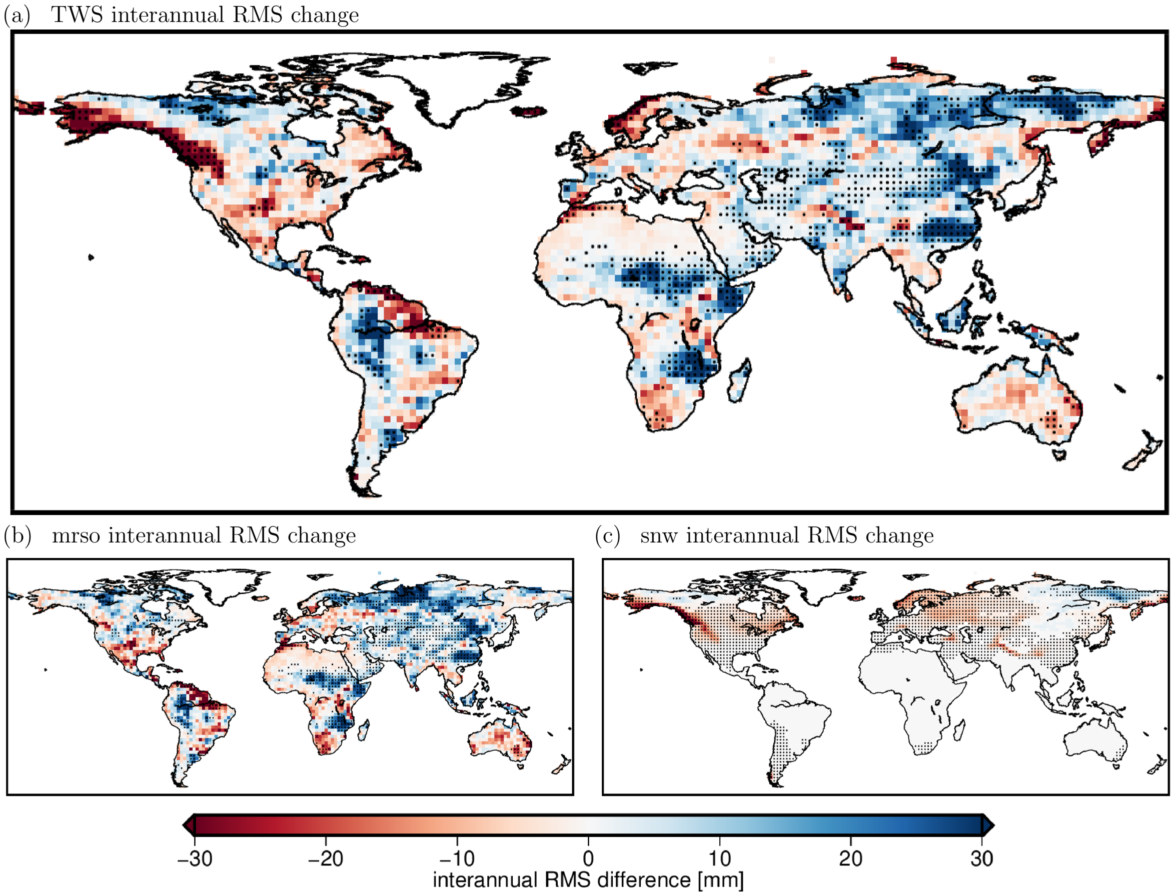

(a) TWS interannual RMS change
(b) mrso interannual RMS change
(c) snw interannual RMS change
−30
−20
−10
0
10
20
30
interannual RMS difference [mm]

Figure 9. (**a**) MMMed mTWS interannual RMS change from 2002–2020 to 2082–2100. Stippling indicates regions where 13 or more of 17 models (i.e., ≥76%) agree on the sign of the RMS change. (**b**,**c**) same as (**a**) but for mrso and snw.

3.2.3. Long-Term Trends

In addition to changes in the annual cycle and interannual variations, the long-term linear trend affects the possible range of future TWS. Centennial mTWS trends in coupled climate models and their relation to GRACE were investigated in Jensen et al. [13] for CMIP5, and we thus give a short update on the findings for CMIP6. The centennial median mTWS linear trend (parameter b in Equation (5)) from each 17 CMIP6 and CMIP5 models is shown in Figure 10. As before, stippling indicates regions of high consensus (≥13 of 17 models agree in sign).

The model consensus among the CMIP6 models is substantially higher than for the CMIP5 models. For CMIP6, in 47% (27% drying, 20% wetting) of the land area more than 76% of the models agree on the trend sign for CMIP6, whereas this number is only 35% (21% drying, 14% wetting) for CMIP5. Especially in Central Africa, South America and Russia/Central Asia the regions of consensus grow from CMIP5 to CMIP6. Furthermore, the median positive and negative trends (0.42 mm/yr and −0.42 mm/yr) are larger for CMIP6 than for CMIP5 (0.18 mm/yr and −0.36 mm/yr). The trend pattern is similar, but it intensifies from CMIP5 to CMIP6 in many regions, e.g., in Central Africa (more intense wetting), tropical South America, and the Mediterranean Coast (more intense drying). In Jensen et al. [13] and Scanlon et al. [37] a general underestimation of the magnitude of linear mTWS trends compared to observations was found. Hence, we conclude that representation of trends improved with the new model generation.

Conclusions on the reliability of centennial linear trends in ESMs are difficult to draw, as the ability of models in representing current trends cannot be verified by observations yet due to the length of the record. Interannual variability largely superimposes possible long-term trends in time spans of less than 30 years, and thus trends over shorter periods are not directly comparable for observations and models [13]. Even if a sufficiently long observational record was available, in some regions trends from observations and models would still not be comparable due to missing processes and components in the models (e.g., groundwater processes, surface water dynamics, anthropogenic contributions).

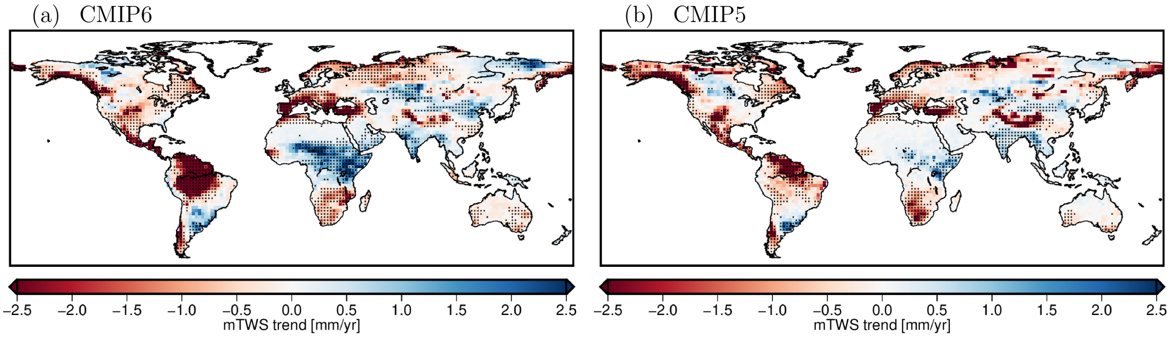


Figure 10. (**a**) MMMed mTWS linear trend over 2000–2100. Stippling indicates regions where 13 or more of 17 models (i.e., ≥76%) agree on the sign of the trend. (**b**) same as (**a**) but for 17 CMIP5 models.

3.3. Detectability of Annual Cycle Changes

Model analysis in Sections 3.2.1 and 3.2.2 revealed that significant long-term changes of TWS variability are to be expected, likely with implications on the future water cycle. Now we investigate to which extent changes as projected with CMIP6 models will be detectable with gravity missions such as GRACE/GRACE-FO or a NGGM. For a first assessment, we assume an observation period of 30 years and restrict the analysis to changes of the annual cycle (amplitude and phase). We consider a signal (i.e., an amplitude or phase change) to be detectable if it exceeds the accuracy of the respective observations. The signal, i.e., the absolute amplitude change for 30 years, is obtained by multiplying the MMMed amplitude change pattern (Figure 7a, given in mm/yr) by 30 years. The absolute phase change for 30 years is derived by scaling the MMMed phase change pattern (Figure 8a; given as absolute change over 100 years) with 0.3 (for 30 years). We then compute accuracies for these absolute 30-yr amplitude and phase changes for two cases, assuming (1) the current GRACE accuracy and (2) a possible NGGM accuracy being 5 times better than for GRACE. This accuracy is proposed as a minimum target performance for NGGMs and is supposed to be achieved by, e.g., employing multiple satellite pairs (double-pair mission), a more favorable orbit design, and improved instrumentation regarding accelerometers and inter-satellite ranging [18,38]. The amplitude and phase change accuracies for cases (1) and (2) are both taken from the accuracies of parameters c' and d' of Equation (5) as described in Section 2.3 under the assumption of a 30 year long time series.

The resulting absolute amplitude and phase change standard deviations (Figure 11a,b) are compared to the absolute amplitude and phase change patterns. In grid cells where the magnitude of the latter exceeds the former the signal is considered to be detectable after 30 years of observations. With a system maintaining the current accuracy of the GRACE mission, and supposing a change pattern as predicted by the MMMed of the 17 CMIP6 models, amplitude changes will be detectable after 30 years in 34% of the land area, and phase changes in 28% of the land area (Figure 11c,d). The threshold for a change being detectable varies spatially as the accuracy pattern is not uniform due to the GRACE error structure. Assuming a NGGM accuracy outperforming the GRACE accuracy by a factor of five but maintaining the same TWS error pattern, amplitude changes would be detectable in 75%, and phase changes in 66% of the land area (Figure 11e,f). Only in very dry regions possible tiny changes in the practically non-existent annual TWS cycle would not be detectable with such a NGGM mission, implying that such changes could be detected reliably everywhere on the continents where agricultural activities take place.

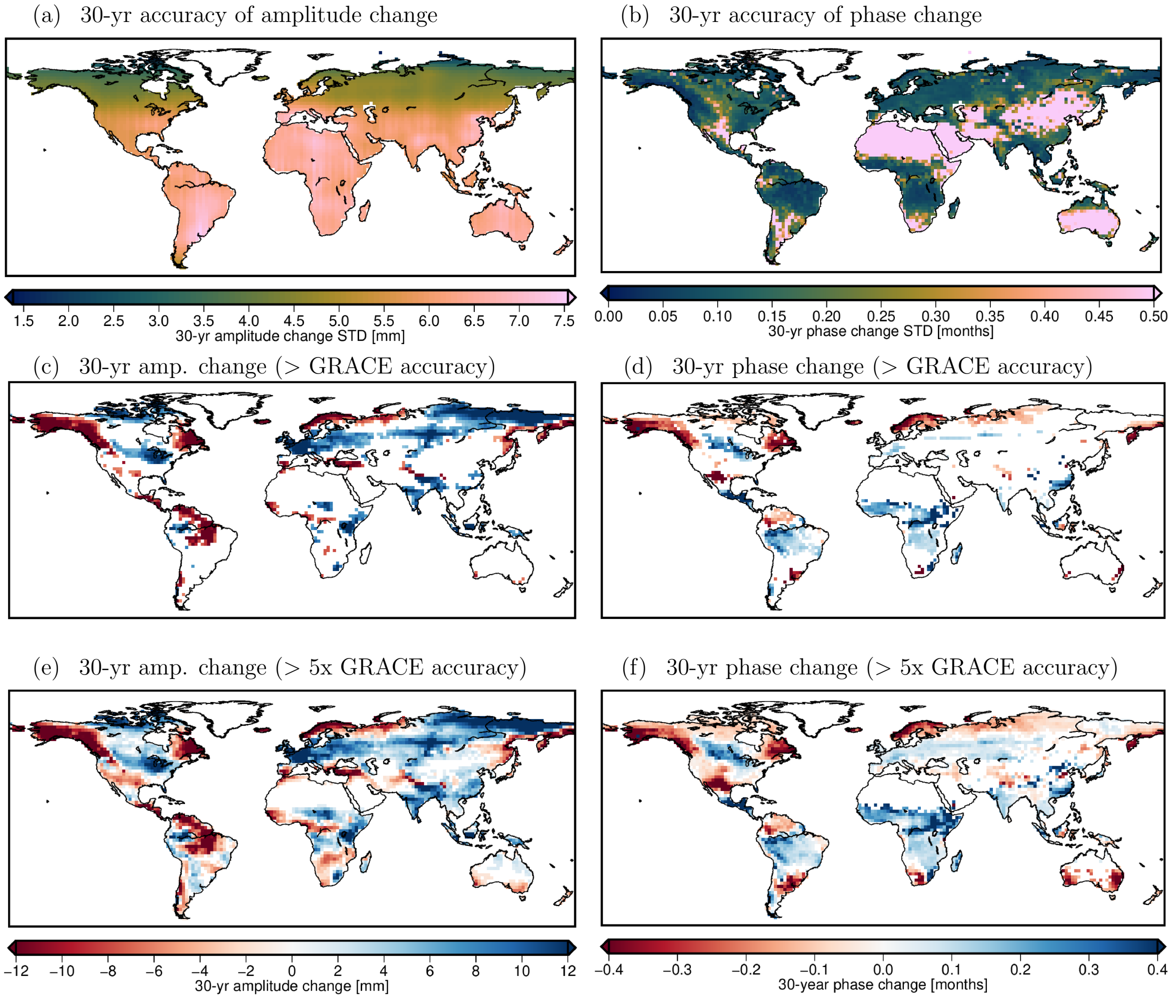

(a) 30-yr accuracy of amplitude change
(b) 30-yr accuracy of phase change
1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5
30-yr amplitude change STD [mm]
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50
30-yr phase change STD [months]
(c) 30-yr amp. change (> GRACE accuracy)
(d) 30-yr phase change (> GRACE accuracy)
(e) 30-yr amp. change (> 5x GRACE accuracy)
(f) 30-yr phase change (> 5x GRACE accuracy)
−12 −10 −8 −6 −4 −2 0 2 4 6 8 10 12
30-yr amplitude change [mm]
−0.4 −0.3 −0.2 −0.1 0.0 0.1 0.2 0.3 0.4
30-year phase change [months]

Figure 11. (**a**) standard deviation of GRACE TWS annual amplitude change over 30 years. (**b**) standard deviation of GRACE TWS phase change of annual cycle over 30 years. (**c**) MMMed mTWS annual amplitude change over 30 years that exceeds the GRACE accuracy (given in (**a**)). (**d**) same as (**c**) but for phase change. (**e**,**f**) same as (**c**,**d**) but assuming the standard deviation of GRACE (given in (**a**,**b**)) being five times smaller.

3.4. Selection of a Representative Model Run for NGGM Simulations

Important prerequisites for the implementation of a NGGM are on the one hand the demonstration of user needs, and on the other hand the justification of science return and societal benefit of a proposed mission concept. Therefore, numerical full-scale simulations are of utmost importance to quantify the achievable performance of an NGGM, and to demonstrate that the science requirements can be met by a certain mission concept and instrumentation. The need for sustained observation of mass transport from space was expressed by an international expert panel under the umbrella of IUGG representing all relevant geoscientific applications [18], and amplified by a resolution adopted by the Council of the International Union of Geodesy and Geophysics [39]. Several numerical simulation studies showed the added value of double-pair concepts for hydrological applications (e.g., [40], and references therein). The main outcome of these numerical simulations is the rating of achievable performance of a mission concept regarding accuracy, spatial and temporal resolution, against the characteristics of the target TWS signal. Up to now, mainly the short-term behaviour of NGGM concepts was evaluated, but hardly any long-term simulation studies focussing on interannual variations and trend signals exist. An exception might be the NGGM study on earthquake detectability by [41], where a mission lifetime of 12 years was simulated to capture full earthquake cycles. For a realistic assessment of the achievable performance over long time periods, it is not sufficient to simply propagate the errors obtained from a short-term simulation to a longer time span. Full-fledged long-term simulations are needed because (1) the relative error contribution of instrument errors and temporal aliasing errors to the total error budget significantly changes with increasing averaging period, (2) they enable a direct parametrization of (linear or non-linear) trends thereby providing a more robust estimate, and (3) they allow for the co-estimation of ocean tides and separation of tidal constituents with very similar excitation periods. While the assessment of the achievable long-term mission performance of potential NGGMs together with an adequate uncertainty characterization is beyond the scope of this study, such simulations need input information on mass changes. Especially for the investigation of climate-driven effects, realistic time series of a possible future development of TWS as input for long-term satellite simulations are of great importance.

In the previous sections the TWS variability was discussed for the (scaled) multi-model median as a best estimate from 17 CMIP6 models with 105 ensemble members in total. However, as elaborated in Section 2.3, the TWS time series of a MMMed is not suitable as input for a NGGM simulation study because the interannual and sub-seasonal variability in the individual model runs is not maintained in the MMMed but largely averaged out. Consequently, for a NGGM simulation study input, a specific model run from the 105 ensemble members has to be selected. Ideally, such a specific model run should be (1) similar to the observations during the GRACE time span, and (2) representative for the expected changes until 2100. This means that for (1) we compare the annual cycle and interannual RMS maps of all 105 ensemble members to the GRACE annual cycle and interannual RMS maps. For (2), we relate the amplitude, phase, and interannual RMS change maps as well as the long-term linear trend maps of the 105 ensemble members to the respective maps of MMMed, which are our best guesses for future TWS variability changes. We recall that MMMed maps are calculated by first decomposing all ensemble members separately and then computing the MMMed for each measure which is different from the decomposition of the MMMed mTWS time series. For the identification of a representative model run we consider two measures to judge similarity between two grids: (a) the pattern correlation to account for the spatial pattern and (b) the RMSD of the empirical cumulative distribution functions (ECDFs) to account for the range of values. During computation of (a) and (b) for the annual cycle we switched from the amplitude and phase representation to coefficients for in-phase and quadrature phase components. This transformation is necessary since only the latter can be assumed to be normally distributed so that conventional statistical metrics can be readily applied. As a result from calculating (a) and (b), for each ensemble member we obtain 14 numbers describing the similarity of the model run with the GRACE observations and with the MMMed: the pattern correlation (a) and ECDF RMSD (b) for (1) sine amplitude, cosine amplitude, and interannual RMS compared to GRACE, and (2) for amplitude change, phase change, interannual RMS change, and linear trend compared to the MMMed.

After computing the 14 metrices for all 105 ensemble members, we perform a ranking: For each metric its range of values is distributed into 100 equidistant classes and dependent on the class into which the respective value of a ensemble member falls, a rank is assigned (1 = smallest rank, smallest similarity; 100 = highest rank, highest similarity). Please note that for the RMSDs of the ECDFs we take the negative values for ranking in order to maintain the rating of the classes (high for good match and low for poor match). Afterward, the ensemble members are sorted with descending mean rank over the 14 metrices. The ranking reveals that no single model run is clearly superior to all others. Instead, model runs of the GFDL-CM4, EC-Earth3, MIRCO6, and CanESM5 exhibit a similar mean performance (Figure 12). Furthermore, the high-ranked model runs do not in every case perform best for all metrices. E.g., the pattern correlation of the annual cycle from most of the top-ranked models with the GRACE annual cycle (first and third column in Figure 12) is only ranked medium high and does not differ substantially from many low-ranked models. On the other hand, an overall low-ranked model does not necessarily occupy poor ranks in all metrices. We note that the rank classification is relative to the range of values of the respective metric but not to its distribution. This means that it is sensitive to outliers: for example, one single model run with an extraordinary high similarity and all others lower (but homogeneous) would lead to the assignment of the first rank to the earlier and the last rank to all later, regardless if the absolute value of the similarity for the other ensemble members may still be pretty good. Thus, in the ranking we do not rate the absolute goodness of similarity but only the performance of the model runs with respect to the others within a specific metric. The full table with the ranking for all 105 ensemble members is given in the Supplementary Material (Figure S6).

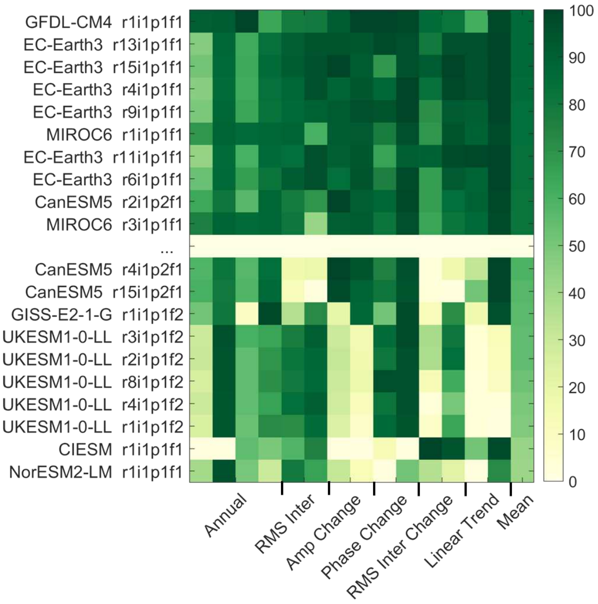


Figure 12. Ranking of the ensemble members according to the classes assigned with pattern correlation (odd columns) and RMSD of ECDF (even columns) of annual cycle and interannual RMS with GRACE (colums 1–6) and amplitude change, phase change, interannual RMS change and linear trend with the MMMed (columds 7–14). Mean rank given in column 15. The upper 10 ensemble members are the best performing, the lower 10 the worst. The full ranking table is provided in the Supplementary Material, Figure S6.

According to the ranking, the model run GFDL-CM4 ri1p1f1 has the best mean match with GRACE and the MMMed on the basis of our selected metrices and rating strategy. Thus, this model run can be considered to be a representative model run of the ensemble regarding its similarity to observations and its alignment with the MMMed and can serve as input for NGGM simulation studies. The mTWS variability in the GFDL-CM4 r1i1p1f1 in terms of annual cycle and interannual variability (with their projected changes) and the long-term linear trend is given in Figure 13. In the Supplementary Materials (Section S8) we provide the results for the detectability of GFDL-CM4 ri1p1f1 annual cycle changes with the GRACE mission or a potential NGGM as described in Section 3.3.

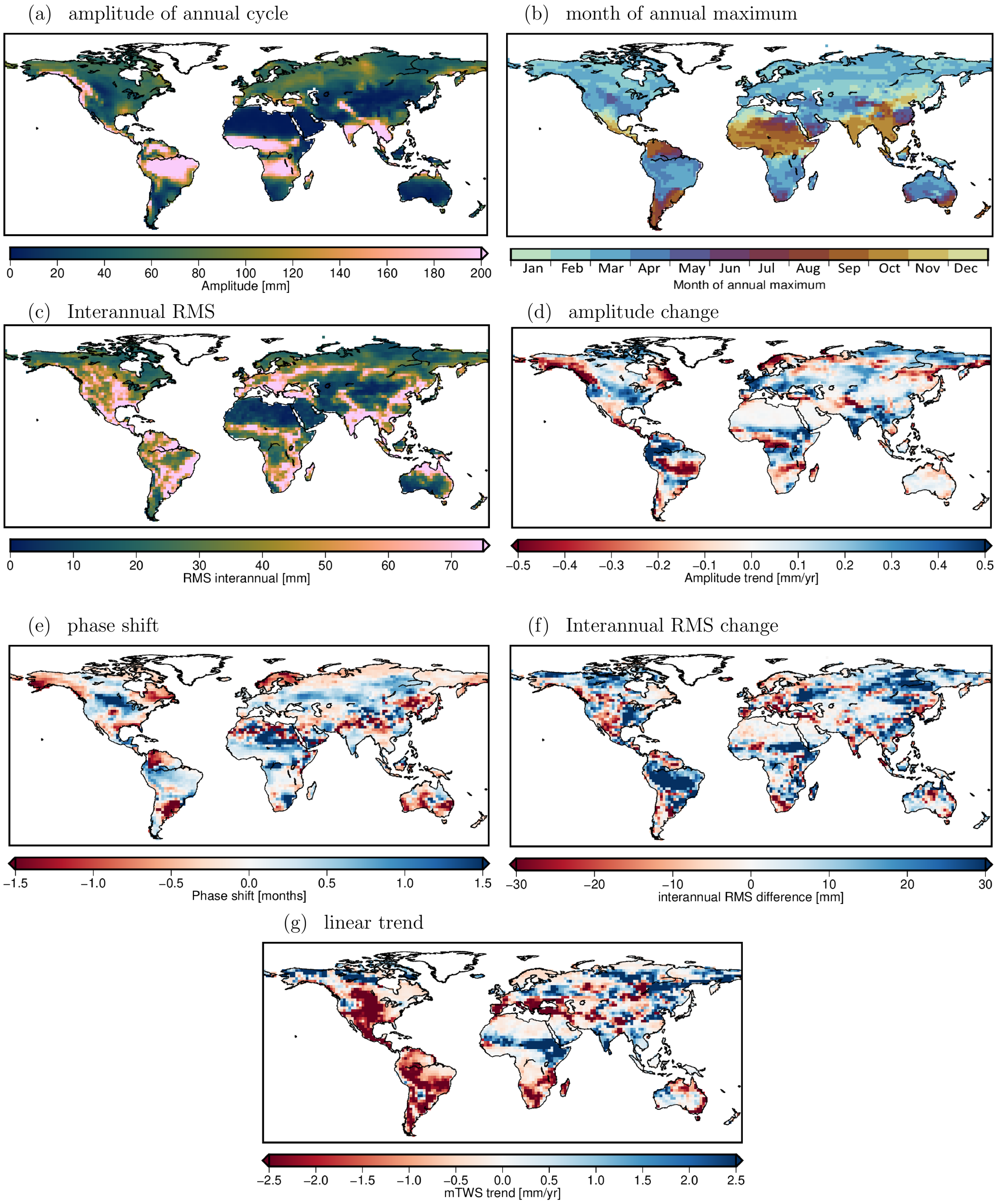


Figure 13. mTWS variability of selected model run GFDL-CM4 r1i1p1f1. (**a**) Amplitude of annual cycle, (**b**) month of the maximum of the annual cycle, (**c**) interannual RMS; (**a**–**c**) for the time span 2002/04–2020/04. (**d**) change of the annual amplitude over 2000–2100, (**e**) phase shift of the annual cycle over 2000–2100, (**f**) interannual RMS change from 2002–2020 to 2082–2100, (**g**) linear trend over 2000–2100.

4. Conclusions

Based on the CMIP6 multi-model ensemble, we assess the current variability of TWS with respect to the observational record acquired by the GRACE and GRACE-FO satellite missions in the time span

2002/04–2020/04 and investigate potential changes in that variability that can be expected to happen until the end of the present century under increasingly changing climate conditions.

In a first step, we compared the general composition of the variances from the seasonal, long-term and sub-seasonal TWS signal in CMIP6 models to the observed signal variances, concluding that the model ensemble represents the current climate conditions reasonably well. While globally the models have a tendency to overestimate the seasonal cycle component and to slightly underestimate the sub-seasonal and long-term signal variance, in equatorial regions we report an overall remarkably good match of all variance components.

To further investigate the fit of different TWS signal components from models and observations we compared the CMIP6 MMMed maps of the annual cycle (amplitude and phase) and the RMS of the interannual signal to the respective GRACE-derived maps. For the annual amplitude the global patterns are similar, with an underestimation by the models in the equatorial climate zone, and an overestimation in the polar zone. Furthermore, the fit is degraded in regions with a small model signal-to-noise-ratio (SNR), mainly in regions with an insignificant annual cycle. For the phase of the annual cycle we found that the models precede the GRACE observations in the majority of the land area (72%) by about half a month in average. We attribute this to missing groundwater processes in CMIP6 models causing an earlier reach of maximum storage due to reduced soil water residence times. The positive time lag of the observations is more pronounced in polar regions than in the other climate zones. For the RMS of the interannual signal the similarities with observations are not as strong as for the annual cycle, and the intermodel spread is larger. As expected from the analysis of the variance components land areas where models underestimate the interannual signal regarding GRACE (60%) exceed areas of overestimation.

In addition to the present-time match of modeled and observed TWS variability, it is important to analyze future changes of TWS variability until 2100. Such information can be useful to refine user requirements for NGGMs that can be readily applied in satellite simulation studies. According to the CMIP6 models, changes in the annual amplitude of regionally up to 27 mm per decade are to be expected. In many regions (45% of the global land area) more than 76% of the models show the same direction of the amplitude change, which is positive in the majority of the land area (56%). The models are less concordant about phase shifts of the annual cycle. In 37% of the land area 76% or more of the models agree on the direction of the shift. A particularly strong phase shift was found for equatorial regions, where until 2100 in 75% of the area the maximum of the annual cycle is projected to be reached on average over two weeks later. The model consensus on changes of the interannual signal is still smaller than for the phase shifts; only in 23% of the land area more than 76% of the models agree on the change direction. Furthermore, the change pattern is more patchy than for the annual cycle with a tendency to an increased interannual variability (54% of the land area exhibits positive changes).

We also made a first step to assess the principle detectability of future changes in the annual cycle of TWS with satellite gravimetry missions. By comparing the amplitude and phase change accuracy achievable with a GRACE-like mission over 30 years with the respective projected changes identified from the CMIP6 models, we derived the regions where changes would be detectable with a mission maintaining the current GRACE accuracy. When anticipating a mission accuracy five times higher than GRACE (which would be a minimum target for a NGGM), changes in the annual cycle in almost all regions with agricultural activities could be detected: whereas with a GRACE-like accuracy only in 34% (amplitude) and 28% (phase) of the land area changes would be observable after 30 years, with a mission of five times higher accuracy, these values would increase to 75% and 66%, where the missing areas are almost exclusively highly arid regions with barely any TWS variability.

To provide a particular time series of TWS changes until 2100 as input for a NGGM satellite simulation study, a representative model run was selected from the CMIP6 multi-model ensemble. In the selection process we considered (1) the similarity of current TWS variability to GRACE observations and (2) the similarity of projected changes to the MMMed change of the respective component. In the ranking of the ensemble members we identified the GFDL-CM4 r1i1p1f1 model run as the most representative one regarding its mean fit to observations and the MMMed change. However, the ranking also revealed that no single model run is clearly superior to all others, but that several model runs exhibit a similar mean fit. The selected model run can serve as a realistic basis for upcoming full-scale simulation studies that include instrument errors, orbit drift, and other systematic errors. Such decades-long satellite simulations with the most promising constellation concepts will demonstrate the added value of such NGGMs for the monitoring of long-term TWS variability.

Supplementary Materials: The following are available online at http://www.mdpi.com/2072-4292/12/23/3898/s1. Figure S1: Regions influenced by surface water, groundwater storage, or glaciers, Figure S2: Identification of independent CMIP6 models, Figure S3: Example for GRACE TWS time series decomposition, Tables S1–S7: Statistics for amplitude, phase, interannual RMS, their changes, and the linear trend in different climate zones, Figure S4: Illustration of phase averaging, Figure S5: Illustration of interference of mrso and snw phases, Figure S6: Ranking of ensemble members, Figure S7: Detectability of annual cycle changes for a specific CMIP6 model run.

Author Contributions: Conceptualization, L.J., A.E., H.D. and R.P.; methodology, L.J., A.E. and H.D.; software, L.J.; validation, L.J.; formal analysis, L.J.; investigation, L.J. and A.E.; resources, A.E.; data curation, L.J.; writing—original draft preparation, L.J.; writing—review and editing, A.E., H.D. and R.P.; visualization, L.J.; supervision, A.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Figure S2 in the Supplementary Material) for producing and making available their model output (e.g., at https://esgf-data.dkrz.de/search/cmip6-dkrz/). For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. Furthermore, we thank Torsten Mayer-Gürr and his group at TU Graz, Austria, for providing the ITSG-Grace2018 data at this website (https://www.tugraz.at/institute/ifg/downloads/gravity-field-models/itsg-grace2018/). The degrees 1 and 2 harmonic coefficients used during GRACE processing were obtained from this site (https://grace.jpl.nasa.gov/data/get-data/).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Eicker, A.; Forootan, E.; Springer, A.; Longuevergne, L.; Kusche, J. Does GRACE see the terrestrial water cycle "intensifying"? *J. Geophys. Res. Atmos.* **2016**, *121*, 2015JD023808. [CrossRef]
2. Greve, P.; Orlowsky, B.; Mueller, B.; Sheffield, J.; Reichstein, M.; Seneviratne, S.I. Global assessment of trends in wetting and drying over land. *Nat. Geosci.* **2014**, *7*, 716. [CrossRef]
3. Konapala, G.; Mishra, A.K.; Wada, Y.; Mann, M.E. Climate change will affect global water availability through compounding changes in seasonal precipitation and evaporation. *Nat. Commun.* **2020**, *11*, 3044. [CrossRef] [PubMed]
4. Kusche, J.; Eicker, A.; Forootan, E.; Springer, A.; Longuevergne, L. Mapping probabilities of extreme continental water storage changes from space gravimetry. *Geophys. Res. Lett.* **2016**, *43*, 8026–8034. [CrossRef]
5. Li, B.; Rodell, M.; Sheffield, J.; Wood, E.; Sutanudjaja, E. Long-term, non-anthropogenic groundwater storage changes simulated by three global-scale hydrological models. *Sci. Rep.* **2019**, *9*, 10746. [CrossRef]
6. Tapley, B.D.; Bettadpur, S.; Watkins, M.; Reigber, C. The gravity recovery and climate experiment: Mission overview and early results. *Geophys. Res. Lett.* **2004**, *31*, L09607. [CrossRef]
7. Kornfeld, R.P.; Arnold, B.W.; Gross, M.A.; Dahya, N.T.; Klipstein, W.M.; Gath, P.F.; Bettadpur, S. GRACE-FO: The Gravity Recovery and Climate Experiment Follow-On Mission. *J. Spacecr. Rocket.* **2019**, *56*, 931–951. [CrossRef]
8. Landerer, F.W.; Flechtner, F.M.; Save, H.; Webb, F.H.; Bandikova, T.; Bertiger, W.I.; Bettadpur, S.V.; Byun, S.H.; Dahle, C.; Dobslaw, H.; et al. Extending the Global Mass Change Data Record: GRACE Follow-On Instrument and Science Data Performance. *Geophys. Res. Lett.* **2020**, *47*, e2020GL088306. . [CrossRef]
9. Tapley, B.D.; Watkins, M.M.; Flechtner, F.; Reigber, C.; Bettadpur, S.; Rodell, M.; Sasgen, I.; Famiglietti, J.S.; Landerer, F.W.; Chambers, D.P.; et al. Contributions of GRACE to understanding climate change. *Nat. Clim. Chang.* **2019**. [CrossRef]
10. Sasgen, I.; Dobslaw, H.; Martinec, Z.; Thomas, M. Satellite gravimetry observation of Antarctic snow accumulation related to ENSO. *Earth Planet. Sci. Lett.* **2010**, *299*, 352–358. [CrossRef]

11. Fasullo, J.T.; Boening, C.; Landerer, F.W.; Nerem, R.S. Australia's unique influence on global sea level in 2010–2011: AUSTRALIA'S INFLUENCE ON 2011 SEA LEVEL. *Geophys. Res. Lett.* **2013**, *40*, 4368–4373. [CrossRef]
12. Rodell, M.; Famiglietti, J.S.; Wiese, D.N.; Reager, J.T.; Beaudoing, H.K.; Landerer, F.W.; Lo, M.H. Emerging trends in global freshwater availability. *Nature* **2018**, *557*, 651. [CrossRef] [PubMed]
13. Jensen, L.; Eicker, A.; Dobslaw, H.; Stacke, T.; Humphrey, V. Long-Term Wetting and Drying Trends in Land Water Storage Derived From GRACE and CMIP5 Models. *J. Geophys. Res. Atmos.* **2019**, *124*, 9808–9823. [CrossRef]
14. Murböck, M.; Pail, R.; Daras, I.; Gruber, T. Optimal orbits for temporal gravity recovery regarding temporal aliasing. *J. Geod.* **2014**, *88*, 113–126. [CrossRef]
15. Bender, P.; Wiese, D.; Nerem, R.S. A possible Dual-GRACE mission with 90 degree and 63 degree inclination orbits. In Proceedings of the 3rd International Symposium on Formation Flying, Missions and Technologies, Noordwijk, The Netherlands, 23–25 April 2008; pp. 59–64.
16. Wiese, D.N.; Visser, P.; Nerem, R.S. Estimating low resolution gravity fields at short time intervals to reduce temporal aliasing errors. *Adv. Space Res.* **2011**, *48*, 1094–1107. [CrossRef]
17. Daras, I.; Pail, R. Treatment of temporal aliasing effects in the context of next generation satellite gravimetry missions. *J. Geophys. Res. Solid Earth* **2017**, *122*, 7343–7362. [CrossRef]
18. Pail, R.; Bingham, R.; Braitenberg, C.; Dobslaw, H.; Eicker, A.; Güntner, A.; Horwath, M.; Ivins, E.; Longuevergne, L.; Panet, I.; et al. Science and User Needs for Observing Global Mass Transport to Understand Global Change and to Benefit Society. *Surv. Geophys.* **2015**, *36*, 743–772. [CrossRef]
19. Flechtner, F.; Neumayer, K.H.; Dahle, C.; Dobslaw, H.; Fagiolini, E.; Raimondo, J.C.; Güntner, A. What Can be Expected from the GRACE-FO Laser Ranging Interferometer for Earth Science Applications? *Surv. Geophys.* **2016**, *37*, 453–470. [CrossRef]
20. Eyring, V.; Bony, S.; Meehl, G.A.; Senior, C.A.; Stevens, B.; Stouffer, R.J.; Taylor, K.E. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.* **2016**, *9*, 1937–1958. [CrossRef]
21. Kvas, A.; Behzadpour, S.; Ellmer, M.; Klinger, B.; Strasser, S.; Zehentner, N.; Mayer-Gürr, T. ITSG-Grace2018: Overview and Evaluation of a New GRACE-Only Gravity Field Time Series. *J. Geophys. Res. Solid Earth* **2019**, *124*, 9332–9344. [CrossRef]
22. Sun, Y.; Riva, R.; Ditmar, P. Optimizing estimates of annual variations and trends in geocenter motion and J_2 from a combination of GRACE data and geophysical models. *J. Geophys. Res. Solid Earth* **2016**, *121*, 8352–8370. [CrossRef]
23. Swenson, S.; Chambers, D.; Wahr, J. Estimating geocenter variations from a combination of GRACE and ocean model output: ESTIMATING GEOCENTER VARIATIONS. *J. Geophys. Res. Solid Earth* **2008**, *113*. [CrossRef]
24. Cheng, M.; Ries, J. The unexpected signal in GRACE estimates of C_{20}. *J. Geod.* **2017**, *91*, 897–914. [CrossRef]
25. Peltier, R.W.; Argus, D.F.; Drummond, R. Comment on "An Assessment of the ICE-6G_C (VM5a) Glacial Isostatic Adjustment Model" by Purcell et al.: The ICE-6G_C (VM5a) GIA model. *J. Geophys. Res. Solid Earth* **2018**, *123*, 2019–2028. [CrossRef]
26. Kusche, J. Approximate decorrelation and non-isotropic smoothing of time-variable GRACE-type gravity field models. *J. Geod.* **2007**, *81*, 733–749. [CrossRef]
27. Lambeck, K. *Geophysical Geodesy: The Slow Deformations of the Earth*; Clarendon Press: New York, NY, USA; Oxford University Press: Oxford, UK, **1988**.
28. Caron, L.; Ivins, E.R.; Larour, E.; Adhikari, S.; Nilsson, J.; Blewitt, G. GIA Model Statistics for GRACE Hydrology, Cryosphere, and Ocean Science. *Geophys. Res. Lett.* **2018**, *45*, 2203–2212. [CrossRef]
29. Han, S.C.; Sauber, J.; Luthcke, S.B.; Ji, C.; Pollitz, F.F. Implications of postseismic gravity change following the great 2004 Sumatra-Andaman earthquake from the regional harmonic analysis of GRACE intersatellite tracking data. *J. Geophys. Res. Solid Earth* **2008**, *113*. [CrossRef]
30. Han, S.C.; Sauber, J.; Luthcke, S. Regional gravity decrease after the 2010 Maule (Chile) earthquake indicates large-scale mass redistribution. *Geophys. Res. Lett.* **2010**, *37*. [CrossRef]
31. Fagiolini, E.; Flechtner, F.; Horwath, M.; Dobslaw, H. Correction of inconsistencies in ECMWF's operational analysis data during de-aliasing of GRACE gravity models. *Geophys. J. Int.* **2015**, *202*, 2150–2158. [CrossRef]

32. Liepert, B.G.; Lo, F. CMIP5 update of 'Inter-model variability and biases of the global water cycle in CMIP3 coupled climate models'. *Environ. Res. Lett.* **2013**, *8*, 029401. [CrossRef]
33. Humphrey, V.; Gudmundsson, L.; Seneviratne, S.I. Assessing Global Water Storage Variability from GRACE: Trends, Seasonal Cycle, Subseasonal Anomalies and Extremes. *Surv. Geophys.* **2016**, *37*, 357–395. [CrossRef] [PubMed]
34. Scanlon, B.R.; Zhang, Z.; Rateb, A.; Sun, A.; Wiese, D.; Save, H.; Beaudoing, H.; Lo, M.H.; Müller-Schmied, H.; Döll, P.; et al. Tracking Seasonal Fluctuations in Land Water Storage Using Global Models and GRACE Satellites. *Geophys. Res. Lett.* **2019**, *46*, 5254–5264. [CrossRef]
35. Peel, M.C.; Finlayson, B.L.; McMahon, T.A. Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci.* **2007**, *11*, 1633–1644. [CrossRef]
36. Dunning, C.M.; Black, E.; Allan, R.P. Later Wet Seasons with More Intense Rainfall over Africa under Future Climate Change. *J. Clim.* **2018**, *31*, 9719–9738. [CrossRef]
37. Scanlon, B.R.; Zhang, Z.; Save, H.; Sun, A.Y.; Müller Schmied, H.; van Beek, L.P.H.; Wiese, D.N.; Wada, Y.; Long, D.; Reedy, R.C.; et al. Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E1080–E1089. [CrossRef]
38. Purkhauser, A.F.; Siemes, C.; Pail, R. Consistent quantification of the impact of key mission design parameters on the performance of next-generation gravity missions. *Geophys. J. Int.* **2020**, *221*, 1190–1210. [CrossRef]
39. IUGG (International Union of Geodesy and Geophysics). Satellite Gravity and Magnetic Mission Constellations. In Proceedings of the Resolutions Adopted by the Council at the XXVI General Assembly, Prague, Czech Republic, 22 June–2 July 2015.
40. Purkhauser, A.F.; Pail, R. Next generation gravity missions: Near-real time gravity field retrieval strategy. *Geophys. J. Int.* **2019**, *217*, 1314–1333. [CrossRef]
41. Cambiotti, G.; Douch, K.; Cesare, S.; Haagmans, R.; Sneeuw, N.; Anselmi, A.; Marotta, A.M.; Sabadini, R. On Earthquake Detectability by the Next-Generation Gravity Mission. *Surv. Geophys.* **2020**, *41*, 1049–1074. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Materials: Emerging Changes in Terrestrial Water Storage Variability as a Target for Future Satellite Gravity Missions

Laura Jensen [1,*], Annette Eicker [1], Henryk Dobslaw [2], and Roland Pail [3]

1. Regions influenced by groundwater, surface water, or glaciers

In some regions TWS from GRACE observations and mTWS from CMIP6 models is not directly comparable since the CMIP6 ESMs do not represent groundwater, surface water storage, and glacier mass changes explicitly. These discrepancies could in principle be reduced by either introducing groundwater, surface water and glacier representations into the ESMs or by removing these components from the integral GRACE signal. The former would be a rewarding effort for future model development as these components have significant impact on land-atmosphere interactions influencing both moisture and energy fluxes and thereby altering climate. The later depends on additional observations or on model results, which introduce uncertainties into the residual GRACE estimates. To identify regions with particularly large differences between TWS and mTWS we compare the full GRACE TWS signal to a TWS signal with the effect of surface water storage and groundwater storage changes removed. In addition, we consider glaciated regions to be affected by discrepancies in TWS and mTWS.

To estimate the effect of surface water storage we make use of an observational data set [1] containing mass change time series of 283 large lakes and reservoirs derived from combining surface water levels (from satellite altimetry) with surface water extent (from remote sensing). By forward modeling these surface water changes to spherical harmonics and applying a spatial DDK3 filtering [2] the data were made comparable to the spatial resolution of GRACE and subsequently mapped to the 2° grid used in this study.

For groundwater long in-situ records are sparse and models exhibit large uncertainties, making a sensible global separation from GRACE-derived TWS unfeasible. However, for natural groundwater variability we consider the discrepancies between models and observations to be minor. Even though the ESMs do not represent groundwater-soil water interactions, they implicitly contain large fractions of the groundwater within their deeper soil layers because the water balance in the models is largely closed and the mass transport to the ocean and atmosphere is limited [3]. Therefore, here we focus on anthropogenic groundwater abstractions only, which are definitely not included in the climate models. We access data from the hydrological model WaterGAP 2.2d [4] where net abstraction groundwater is defined as groundwater withdrawals minus return flow from irrigation. Since irrigation is partly taken from surface waters, net abstraction from groundwater lead to mass increase. We convert the monthly (2003/01 – 2016/12) global grids from fluxes (in mm/s) to monthly accumulated water storage change (EWH in mm) and remap it to 2° spatial resolution. We then apply a GRACE-like spatial filtering (DDK3 filter) [2].

Afterward, we compute the RMS over 2003/01 – 2016/12 of the annual plus interannual signal (the main signal components discussed in the paper) and compare (1) the full GRACE time series and (2) the GRACE time series minus the surface water storage and net abstraction time series. The relative difference of these two RMS values is displayed in Figure S1. The differences are mostly positive, which means that the signal gets smaller after removing estimates of surface water storage and (anthropogenic) groundwater change, which is expected. For surface water storage (Figure S1a) the RMS reduction is generally larger than for the net abstraction (Figure S1b). Smaller values for the net abstraction are due to the fact that they mainly occur as a linear mass trends (which is not investigated here) whereas seasonal and year-to-year variations are minor. The relative RMS difference for the combined surface water and net abstraction effect is displayed in Figure S1c. Regions where the difference exceeds 10% of the total (annual and interannual) signal are shown in red in Figure S1d. These regions make up 8% of the land surface (Greenland and Antarctica excluded).

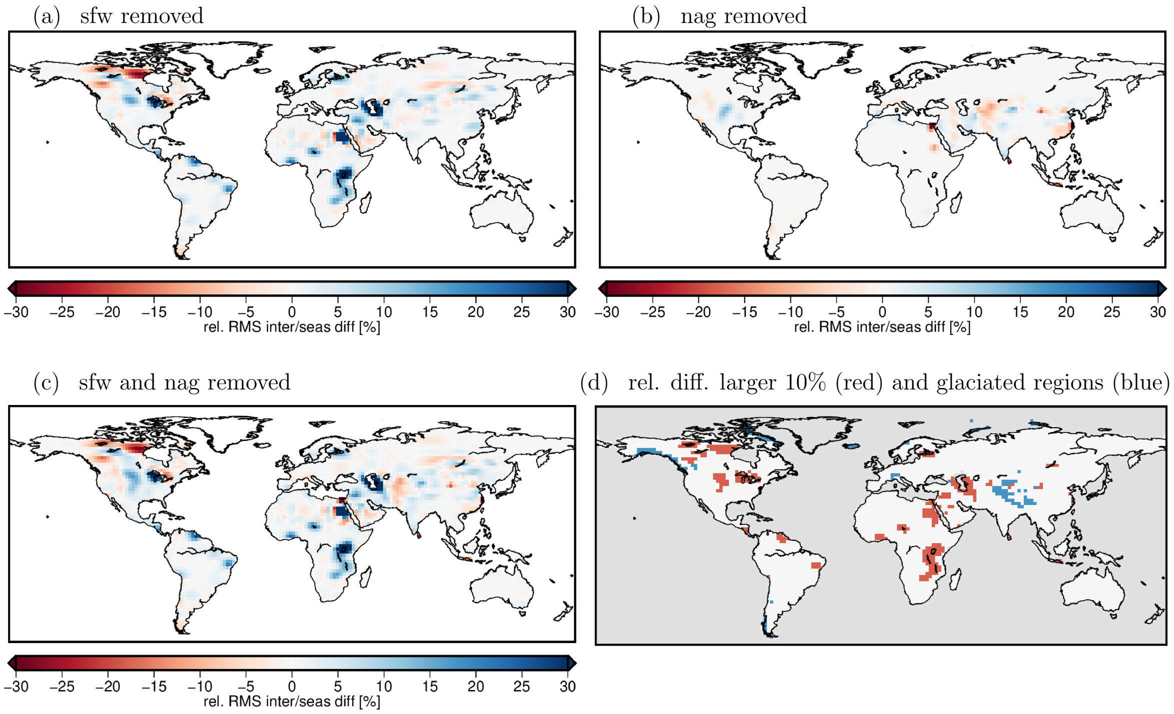


Figure S1. (**a**) Relative difference between the interannual plus annual RMS of GRACE and GRACE reduced by surface water storage changes for the time span 2003 – 2016. (**b**) same as (a) for net groundwater abstraction. (**c**) same as (a) for combined surface water and net abstraction changes. (**d**) regions where relative difference from (c) exceeds 10% (red) and regions partly covered by glaciers (blue).

The reduction of the GRACE time series by the two data sets described above provides a good estimate which regions are most affected by TWS and mTWS discrepancies. However, the currently available data set for surface water changes is still affected by large uncertainties and cannot capture all surface water bodies due to a spatial undersampling of, e.g., rivers and smaller lakes. This can be expected to improve once the SWOT data [5] become available. Also groundwater abstractions from hydrological modeling have uncertainties that are difficult to quantify due to sparse observations. Accurate quantification of both components still requires more research and with respect to the impractical quantification of errors we abstained from actually removing the surface water and abstraction data from the GRACE observations in the actual comparison with CMIP6 models in the main article. We rather indicate in Figure S1d regions where the results have to be interpreted with care as the discrepancies between observed and modeled TWS are probably larger here than in the remaining land area.

In addition to groundwater and surface water, also glacier mass changes are not contained in CMIP6 models, but observed by GRACE. In glaciated regions, moisture dynamics are not purely driven by soil moisture and snow variability but also influenced by ice mass changes, constituting a discrepancy between TWS and mTWS. To identify these regions, we access the GLIMS (Global Land Ice Measurements from Space) Glacier Database [6], currently containing the outlines of about 546300 glaciers from around the world. To quantify the degree to which a 2° grid cell (the resolution of the maps in this study) is covered by glaciers, we raster the GLIMS glacier polygons to a rather fine resolution of 0.025° and afterward count the number of 0.025° glacier grid cells within a 2° grid cell. In Figure S1d all 2° grid cells where more than 100 small 0.025° grid cells are glaciated are marked in blue. These regions make up 3% of the land surface.

2. Identification of independent CMIP6 models

At the time of writing, the CMIP6 data base contains mrso and snw data from 25 models. However, some of these models share the same sub-models (land, ocean, or atmosphere) or are extensions and sub-versions from each other. Thus, not all model results can be considered to be independent from each other. If all 25 models would contribute to the multi-model median, it would be biased towards particular models because their data would be included multiple times. Therefore, we exclude dependent models. Two models are considered not to be independent from each other when their long-term (2000 – 2100) linear trend patterns are very similar. This criterion was also applied in Jensen *et al.* [7] and has been shown to effectively identify models with common components. The trend pattern of each model is obtained from the ensemble mean of all runs belonging to the respective model via least squares adjustment. The similarity between the trend maps is measured by the pattern correlation, which is calculated as Pearson's product-moment correlation coefficient of the vectorized trend maps (excluding ocean grid cells). The pattern correlation of the trend maps between all 25 model ensemble means is displayed in Figure S2. A correlation of 70% or more was defined as the threshold for a model to be excluded. The criteria for the selection of a specific model from the highly correlated models were its degree of specialization (most general), spatial resolution (closest to 2°), or the number of ensemble members (most members).

The models excluded by these criteria are marked in light gray in Figure S2. Detailed information and references for the 17 models remaining (bold black) can be accessed, e.g., via https://esgf-data.dkrz.de/projects/cmip6-dkrz/ (last visit: 9/24/2020).

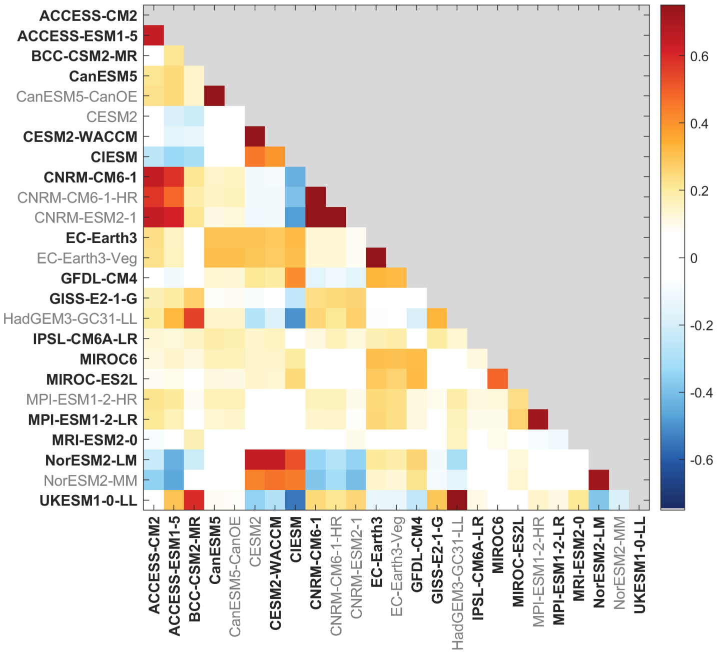


Figure S2. Correlations of long-term (2000 – 2100) linear trend patterns from 25 CMIP6 model ensemble means. Models excluded from the analysis are marked in light gray, models used in the study are highlighted in bold black.

3. Example for signal decomposition of GRACE TWS time series

An example TWS time series from GRACE and its decomposition into long-term, seasonal and sub-seasonal components is shown in Figure S3. Compared to the modeled time series (Figure 1 of main text) the sub-seasonal component is larger in GRACE. This is probably partly due to observational noise, but also due to real signals captured by GRACE on a month-to-month time scale. It has been shown that GRACE can observe TWS signals even on time scales down to a few days [8], and it is not clear to which extent these variations are reproduced by the ESMs. Therefore, we concentrate in our study on seasonal and interannual variations. Please note that the outlier in the GRACE data (January 2015) is due to a repeat-orbit constellation leading to a degraded monthly solution. However, this does only marginally influence the estimation of the seasonal and interannual parameters.

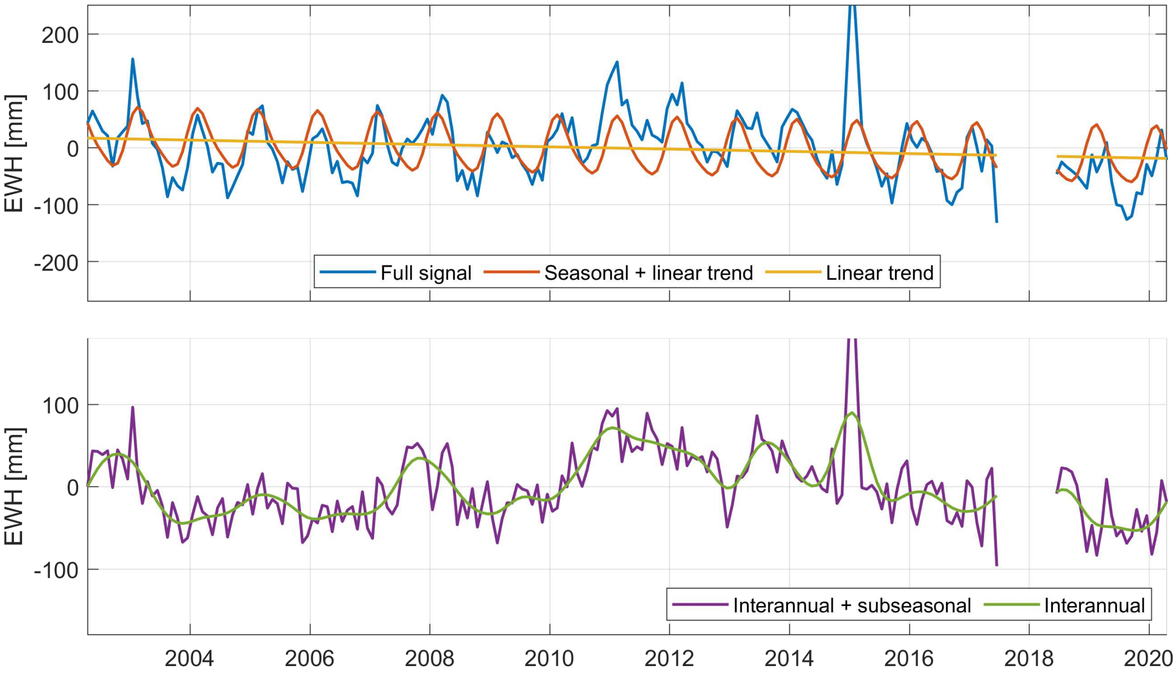


Figure S3. Example for the decomposition of a GRACE TWS time series into linear trend, seasonal, subseasonal and interannual signal. The location is 13°E and 52.5°N (Potsdam, Germany).

4. Statistics for different climate zones

Table S1. Comparison of annual amplitude from GRACE and CMIP6 MMMed (2002/04 – 2020/04) for different climate zones. Column 1: percentage of land area lying in the respective climate zone; column 2: percentage of area with a SNR $<$ 1; column 3: percentage of land area lying in the respective climate zone after excluding areas with SNR $<$ 1; columns 4 and 5: area percentage of overestimation (ratio $>$ 1) and underestimation (ratio $\leq$ 1) of the CMIP6 amplitude w.r.t. the GRACE amplitude (after excluding regions with SNR $<$ 1); columns 6 and 7: median of the ratios in the over- and underestimation areas.

Amplitude	% area	% SNR < 1	% area w/o SNR < 1	% over	% under	median over	median under
global	100.0	24.0	100.0	57.5	42.5	1.38	0.79
equatorial	21.7	10.0	25.8	50.6	49.4	1.46	0.76
arid	36.0	49.5	24.0	55.2	44.8	1.51	0.70
temperate	16.1	10.1	19.0	57.0	43.0	1.41	0.74
polar	26.2	9.9	31.2	65.4	34.6	1.28	0.87

Table S2. As table S1, but for the phase of the annual cycle. Columns 1 and 2: area percentage of positive time shift (models earlier) and negative time shift (models later) of the CMIP6 phase w.r.t. the GRACE phase (after excluding regions with SNR $<$ 1); columns 3 and 4: median of the phase shift in the positive and negative areas.

Phase	% earlier	% later	median earlier [months]	median later [months]
global	71.9	28.1	0.50	-0.32
equatorial	62.1	37.9	0.43	-0.36
arid	72.5	27.5	0.59	-0.48
temperate	68.2	31.8	0.66	-0.27
polar	81.8	18.3	0.45	-0.20

Table S3. As table S1, but for the RMS of the interannual signal. Column 1: percentage of area with a SNR $<$ 1; column 2: percentage of land area lying in the respective climate zone after excluding areas with SNR $<$ 1; columns 3 and 4: area percentage of overestimation (ratio $>$ 1) and underestimation (ratio $\leq$ 1) of the CMIP6 interannual RMS w.r.t. the GRACE interannual RMS (after excluding regions with SNR $<$ 1); columns 5 and 6: median of the ratios in the over- and underestimation areas.

Inter. RMS	% SNR < 1	% area w/o SNR < 1	% over	% under	median over	median under
global	14.7	100.0	40.2	59.9	1.30	0.69
equatorial	12.4	22.6	46.5	53.5	1.39	0.76
arid	29.6	29.6	32.2	67.8	1.38	0.56
temperate	3.6	18.1	42.7	57.3	1.30	0.67
polar	3.1	29.7	41.7	58.3	1.21	0.77

Table S4. Analysis of CMIP6 MMMed annual amplitude changes (2000/01 – 2100/12) for different climate zones. Upper part, column 1: the percentage of land area where > 75% of the models agree on the sign of the trend (high consensus); columns 2 and 3: area percentage of positive and negative trends; columns 4 and 5: median trend calculated separately over the positive and negative areas. Lower part, column 1: percentage of land area lying in the respective climate zone when restricting to regions of high model consensus; columns 2 to 5: as upper part but for high consensus regions only.

Amp. Change	% highcons	% pos	% neg	median pos [mm/yr]	median neg [mm/yr]
global	44.5	56.0	44.0	0.12	-0.11
equatorial	34.0	51.3	48.8	0.13	-0.18
arid	41.9	47.3	52.7	0.06	-0.04
temperate	48.2	61.5	38.5	0.13	-0.15
polar	54.3	68.5	31.5	0.18	-0.21
high consensus	% area				
global	100.0	66.3	33.7	0.21	-0.26
equatorial	16.6	58.6	41.4	0.26	-0.37
arid	33.9	59.9	40.1	0.11	-0.15
temperate	17.4	70.3	29.7	0.20	-0.31
polar	32.0	74.8	25.2	0.27	-0.42

Table S5. As table S4, but for changes of the phase of the annual cycle. Upper part, column 1: the percentage of land area where > 75% of the models agree on the sign of the phase shift (high consensus); columns 2 and 3: area percentage of positive (later) and negative (earlier) phase shifts; columns 4 and 5: median phase shift calculated separately over the positive and negative areas. Lower part, column 1: percentage of land area lying in the respective climate zone when restricting to regions of high model consensus; columns 2 to 5: as upper part but for high consensus regions only.

Phase shift	% highcons	% pos/later	% neg/earlier	median pos [months]	median neg [months]
global	36.7	54.8	45.2	0.39	-0.35
equatorial	43.8	74.8	25.3	0.49	-0.38
arid	32.3	55.3	44.7	0.46	-0.38
temperate	39.4	49.3	50.7	0.35	-0.31
polar	35.1	40.9	59.1	0.25	-0.33
high consensus	% area				
global	100.0	60.7	39.3	0.70	-0.78
equatorial	25.9	85.0	15.0	0.72	-0.77
arid	31.7	66.6	33.4	0.94	-0.95
temperate	17.2	55.6	44.4	0.56	-1.15
polar	25.1	31.8	68.2	0.45	-0.65

Table S6. As table S4, but for changes of the RMS of the interannual signal. Upper part, column 1: the percentage of land area where > 75% of the models agree on the sign of the change (high consensus); columns 2 and 3: area percentage of positive and negative changes; columns 4 and 5: median interannual RMS change calculated separately over the positive and negative areas. Lower part, column 1: percentage of land area lying in the respective climate zone when restricting to regions of high model consensus; columns 2 to 5: as upper part but for high consensus regions only.

Inter. RMS change	% highcons	% pos	% neg	median pos [mm]	median neg [mm]
global	22.6	54.1	45.9	7.02	-5.75
equatorial	14.6	50.0	50.0	8.14	-6.42
arid	31.3	53.9	46.2	5.40	-3.92
temperate	24.8	56.3	43.7	8.02	-7.12
polar	16.1	56.6	43.4	7.98	-6.50
high consensus	% area				
global	0.0	77.8	22.2	13.49	-23.79
equatorial	14.2	75.6	24.4	16.28	-25.94
arid	49.7	82.3	17.7	9.48	-15.31
temperate	17.6	76.3	23.7	19.37	-26.95
polar	18.6	68.6	31.4	18.50	-31.69

Table S7. As table S4, but for the linear trend and for both, CMIP6 and CMIP5. Upper part, column 1: the percentage of land area where > 75% of the models agree on the sign of the change (high consensus); columns 2 and 3: area percentage of positive and negative trends; columns 4 and 5: median trend calculated separately over the positive and negative areas. Lower part, column 1: percentage of land area lying in the respective climate zone when restricting to regions of high model consensus; columns 2 to 5: as upper part but for high consensus regions only.

Trend CMIP6	% highcons	% pos	% neg	median pos [mm/yr]	median neg [mm/yr]
global	47.2	42.9	57.1	0.42	-0.42
equatorial	57.3	49.9	50.1	0.60	-1.73
arid	47.3	54.0	46.0	0.31	-0.30
temperate	45.9	35.6	64.4	0.42	-0.44
polar	39.3	26.5	73.5	0.37	-0.40
high consensus	% area				
global	100.0	30.1	70.0	1.01	-0.89
equatorial	26.4	34.1	65.9	1.13	-2.76
arid	36.2	44.0	56.1	0.90	-0.58
temperate	15.6	15.5	84.5	1.10	-1.00
polar	21.9	12.6	87.4	1.01	-0.71

Trend CMIP5	% highcons	% pos	% neg	median pos [mm/yr]	median neg [mm/yr]
global	34.6	40.8	59.2	0.18	-0.36
equatorial	39.7	46.5	53.5	0.27	-0.54
arid	33.1	53.3	46.7	0.12	-0.27
temperate	43.9	27.3	72.7	0.28	-0.35
polar	26.8	27.4	72.7	0.30	-0.44
high consensus	% area				
global	100.0	24.0	76.1	0.62	-0.91
equatorial	24.4	33.0	67.0	0.62	-1.19
arid	34.6	31.6	68.4	0.51	-0.73
temperate	20.5	14.8	85.2	0.68	-0.80
polar	20.6	9.4	90.6	0.67	-1.02

5. Averaging phases

Due to the fact that phases of the annual cycle are not normally distributed, the calculation of statistical measures like mean, median, or standard deviation is not straight forward. To overcome this, we apply an iterative algorithm that is illustrated in Figure S4. The upper panel shows the histogram of an artificial data set of 13 phase values (in month of the year, i.e., $0, \ldots, 1$ = January, $1, \ldots, 2$ = February, and so on) for which the mean, median, and standard deviation shall be calculated. The arithmetic mean of the phase values is shown as a vertical red line, and corresponding numbers for median, mean and standard deviation are given on the right of the panel. The value 6.8462 (i.e., July) is obviously not the correct mean phase, which should be between December and January. Now we iteratively shift all phase values smaller or equal than k months (with $k = 1, \ldots, 6$) by 12 months and calculate median, mean, and standard deviation again (panels 2 – 4 in Figure S4). We define the mean and median where the standard deviation of the (shifted) sample gets minimal as the best estimates for the statistical measures (panel 4). If the computed mean or median from the shifted sample happens to be larger than 12, we subtract 12, in order to keep the range of values from January to December. In this example, the mean is 0.3077 (i.e., January, vertical yellow line in panel 4), and the median is 12 (i.e. December), with a standard deviation of 1.6525 months.

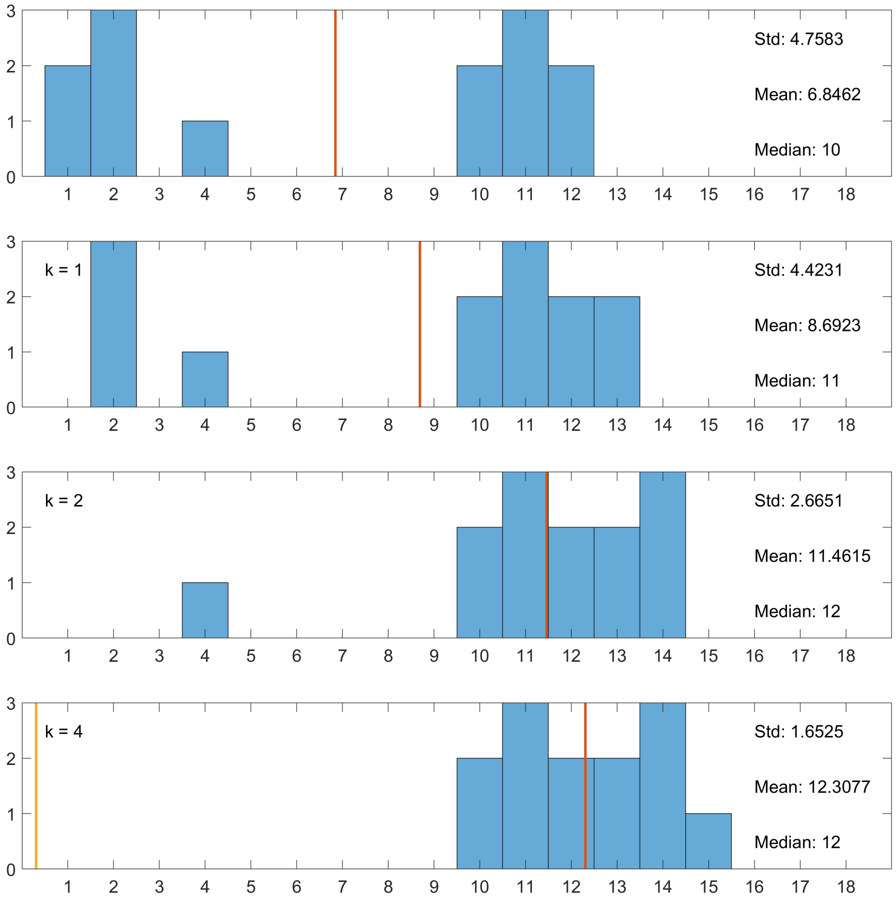


Figure S4. Illustration for the computation of the mean and median phase.

6. Interference of mrso and snw phases

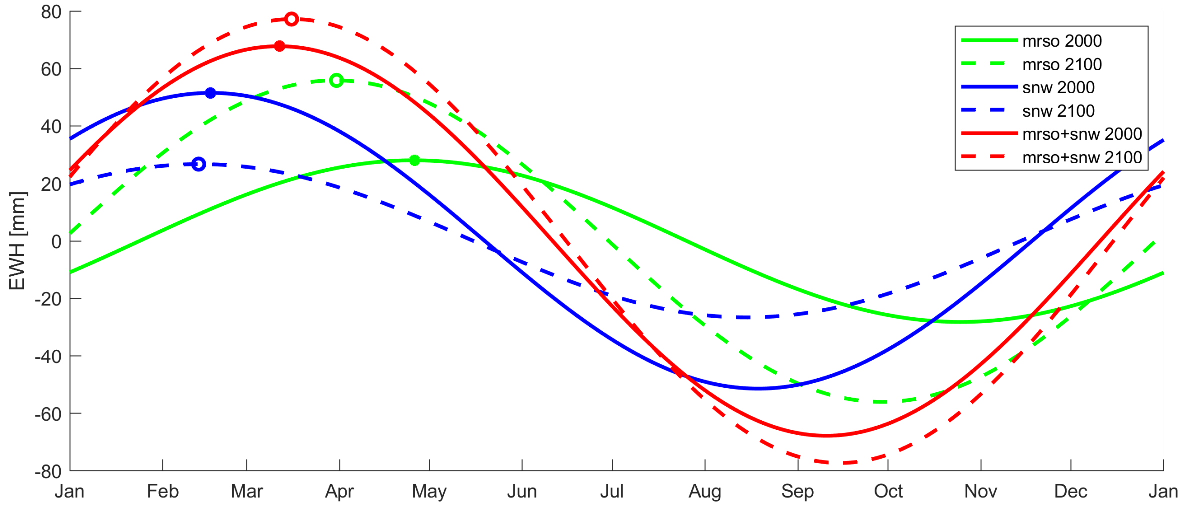


Figure S5. Example for a time series where the time shift from 2000 to 2100 is negative for mrso and snw, but positive for mTWS (i.e. the sum of mrso and snw). The location is 75.0°E and 57.0°N (Russia).

7. Ranking of ensemble members

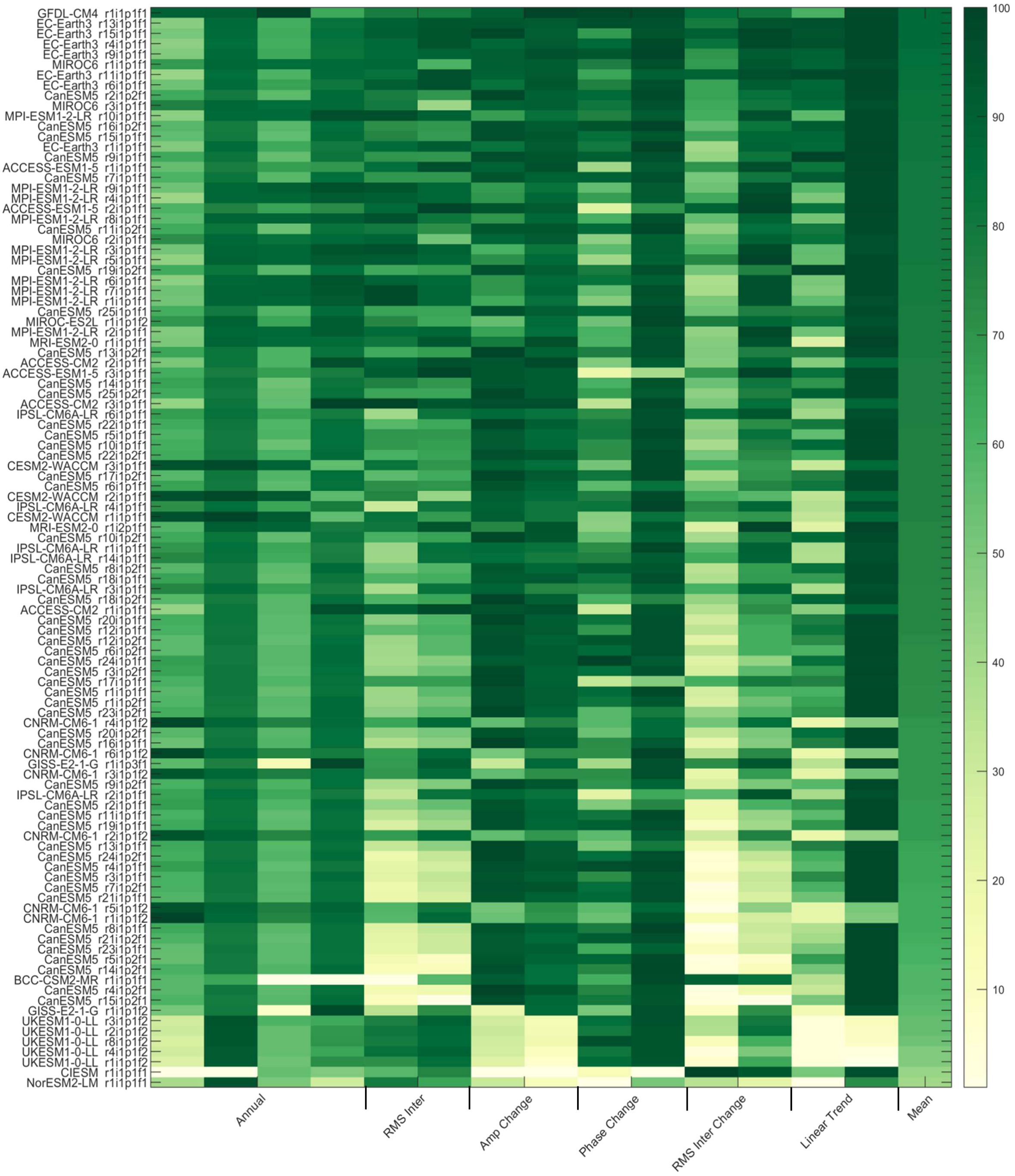


Figure S6. Ranking of the ensemble members according to the classes assigned with pattern correlation (odd columns) and RMSD of ECDF (even columns) of annual cycle and interannual RMS with GRACE (colums 1–6) and amplitude change, phase change, interannual RMS change and linear trend with the MMMed (columns 7–14).

8. Detectability of annual cycle changes for a specific CMIP6 model run

In Figure S7 we compare the amplitude and phase change patterns of the r1i1p1f1 run of the GFDL-CM4 model (selected as a representative model run in Section 3.4 of main text) to (1) the current GRACE accuracy and (2) a five times smaller accuracy of a possible NGGM (cf. Section 3.3 of main text). The area percentages of detectable changes are 40% (amplitude) and 27% (phase) for the GRACE accuracy and 77% (amplitude) and 68% (phase) for the NGGM accuracy.

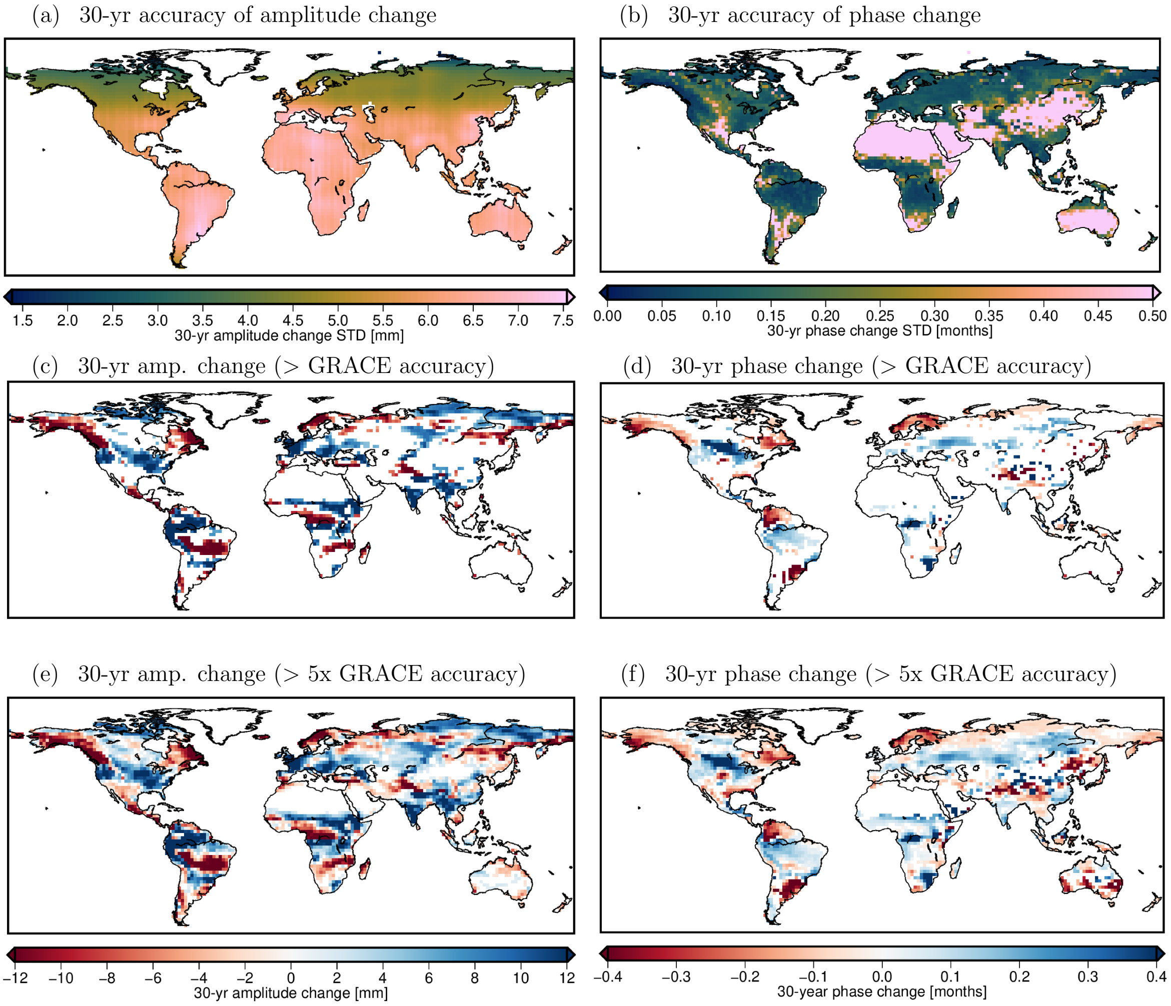

(a) 30-yr accuracy of amplitude change
(b) 30-yr accuracy of phase change
1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.5 6.0 6.5 7.0 7.5
30-yr amplitude change STD [mm]
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50
30-yr phase change STD [months]
(c) 30-yr amp. change (> GRACE accuracy)
(d) 30-yr phase change (> GRACE accuracy)
(e) 30-yr amp. change (> 5x GRACE accuracy)
(f) 30-yr phase change (> 5x GRACE accuracy)
−12 −10 −8 −6 −4 −2 0 2 4 6 8 10 12
30-yr amplitude change [mm]
−0.4 −0.3 −0.2 −0.1 0.0 0.1 0.2 0.3 0.4
30-year phase change [months]

Figure S7. (**a**) standard deviation of GRACE TWS annual amplitude change over 30 years. (**b**) standard deviation of GRACE TWS phase change of annual cycle over 30 years. (**c**) mTWS annual amplitude change of the GFDL-CM4 r1i1p1f1 over 30 years that exceeds the GRACE accuracy (given in (a)). (**d**) same as (c) but for phase change. (**e**,**f**) same as (c,d) but assuming the standard deviation of GRACE (given in (a) and (b)) being five times smaller.

References

1. Deggim, S.; Eicker, A.; Schawohl, L.; Gerdener, H.; Schulze, K.; Engels, O.; Kusche, J.; Saraswati, A.T.; van Dam, T.; Ellenbeck, L.; Dettmering, D.; Schwatke, C.; Mayr, S.; Klein, I.; Longuevergne, L. RECOG RL01: Correcting GRACE total water storage estimates for global lakes/reservoirs and earthquakes. *Earth System Science Data Discussions* **2020**, pp. 1–30. Publisher: Copernicus GmbH, doi:https://doi.org/10.5194/essd-2020-256.
2. Kusche, J. Approximate decorrelation and non-isotropic smoothing of time-variable GRACE-type gravity field models. *Journal of Geodesy* **2007**, *81*, 733–749. doi:10.1007/s00190-007-0143-3.

3. Liepert, B.G.; Lo, F. CMIP5 update of 'Inter-model variability and biases of the global water cycle in CMIP3 coupled climate models'. *Environmental Research Letters* **2013**, *8*, 029401. doi:10.1088/1748-9326/8/2/029401.
4. Müller Schmied, H.; Cáceres, D.; Eisner, S.; Flörke, M.; Herbert, C.; Niemann, C.; Peiris, T.A.; Popat, E.; Portmann, F.T.; Reinecke, R.; Schumacher, M.; Shadkam, S.; Telteu, C.E.; Trautmann, T.; Döll, P. The global water resources and use model WaterGAP v2.2d: Model description and evaluation. *Geoscientific Model Development Discussions* **2020**, pp. 1–69. Publisher: Copernicus GmbH, doi:https://doi.org/10.5194/gmd-2020-225.
5. Alsdorf, D.; Rodriguez, E.; Morrow, R.; Mognard, N.; Lambin, J.; Vaze, P.; Lafon, T. THE SURFACE WATER AND OCEAN TOPOGRAPHY (SWOT) MISSION **2010**.
6. GLIMS.; NSIDC. Global Land Ice Measurements from Space glacier database. Compiled and made available by the international GLIMS community and the National Snow and Ice Data Center, Boulder CO, U.S.A. **2005, updated 2020**. Publisher: NSIDC, doi:10.7265/N5V98602.
7. Jensen, L.; Eicker, A.; Dobslaw, H.; Stacke, T.; Humphrey, V. Long-Term Wetting and Drying Trends in Land Water Storage Derived From GRACE and CMIP5 Models. *Journal of Geophysical Research: Atmospheres* **2019**, *124*, 9808–9823. doi:10.1029/2018JD029989.
8. Eicker, A.; Jensen, L.; Wöhnke, V.; Dobslaw, H.; Kvas, A.; Mayer-Gürr, T.; Dill, R. Daily GRACE satellite data evaluate short-term hydro-meteorological fluxes from global atmospheric reanalyses. *Scientific Reports* **2020**, *10*, 4504. Number: 1 Publisher: Nature Publishing Group, doi:10.1038/s41598-020-61166-0.

A.3. Predictive Skill Assessment for Land Water Storage in CMIP5 Decadal Hindcasts by a Global Reconstruction of GRACE Satellite Data

Reference

Jensen, L., Eicker, A., Stacke, T., Dobslaw, H. (2020). Predictive Skill Assessment for Land Water Storage in CMIP5 Decadal Hindcasts by a Global Reconstruction of GRACE Satellite Data. J. Climate 33, 9497–9509. https://doi.org/10.1175/JCLI-D-20-0042.1

Abstract

The evaluation of decadal climate predictions against observations is crucial for their benefit to stakeholders. While the skill of such forecasts has been verified for several atmospheric variables, land-hydrological states such as terrestrial water storage (TWS) have not been extensively investigated yet due to a lack of long observational records. Anomalies of TWS are globally observed with the satellite missions GRACE (2002 - 2017) and GRACE-FO (since 2018). By means of a GRACE-like reconstruction of TWS available over 41 years, we demonstrate that this data type can be used to evaluate the skill of decadal prediction experiments made available from different Earth System Models as part of both CMIP5 and CMIP6. Analysis of correlation and root-mean-square deviation (RMSD) reveals that for the global land average the initialized simulations outperform the historical experiments in the first three forecast years. This predominance originates mainly from equatorial regions where we assume a longer influence of initialization due to longer soil memory times. Evaluated for individual grid cells, the initialization has a largely positive effect on the forecast year 1 TWS states, however, a general grid-scale prediction skill for TWS of more than two years could not be identified in this study for CMIP5. First results from decadal hindcasts of three CMIP6 models indicate a predictive skill comparable to CMIP5 for the multi-model mean in general, and a distinct positive influence of the improved soil-hydrology scheme implemented in the MPI-ESM for CMIP6 in particular.

Declaration of own contribution

Table A.3.: Contribution to Paper No. 3.

Involved in	Estimated contribution
Ideas and conceptual design	90%
Computation and results	100%
Analysis and interpretation	90%
Manuscript, figures and tables	90%
Total	92,5%

Confirmation of Co-Authors

I hereby confirm the correctness of the declaration of the contribution of Laura Jensen for Paper No. 3 in Table A.3:

26.05.21

Annette Eicker — Date
(HCU Hamburg)

Tobias Stacke — 28.05.21

Tobias Stacke — Date
(HZG Geesthacht)

Henryk
Dobslaw

Digital unterschrieben von
Henryk Dobslaw
Datum: 2021.05.26 20:28:22
+02'00'

Henryk Dobslaw — Date
(GFZ Potsdam)

Predictive Skill Assessment for Land Water Storage in CMIP5 Decadal Hindcasts by a Global Reconstruction of GRACE Satellite Data

LAURA JENSEN AND ANNETTE EICKER

Geodesy and Geoinformatics, HafenCity University, Hamburg, Germany

TOBIAS STACKE

Helmholtz-Zentrum Geesthacht, Centre for Materials and Coastal Research, Geesthacht, Germany

HENRYK DOBSLAW

Helmholtz Centre Potsdam, German Research Centre for Geosciences (GFZ), Potsdam, Germany

(Manuscript received 22 January 2020, in final form 13 August 2020)

ABSTRACT: The evaluation of decadal climate predictions against observations is crucial for their benefit to stakeholders. While the skill of such forecasts has been verified for several atmospheric variables, land hydrological states such as terrestrial water storage (TWS) have not been extensively investigated yet due to a lack of long observational records. Anomalies of TWS are globally observed with the satellite missions GRACE (2002–2017) and GRACE-FO (since 2018). By means of a GRACE-like reconstruction of TWS available over 41 years, we demonstrate that this data type can be used to evaluate the skill of decadal prediction experiments made available from different Earth system models as part of both CMIP5 and CMIP6. Analysis of correlation and root-mean-square deviation (RMSD) reveals that for the global land average the initialized simulations outperform the historical experiments in the first three forecast years. This predominance originates mainly from equatorial regions where we assume a longer influence of initialization due to longer soil memory times. Evaluated for individual grid cells, the initialization has a largely positive effect on the forecast year 1 TWS states; however, a general grid-scale prediction skill for TWS of more than 2 years could not be identified in this study for CMIP5. First results from decadal hindcasts of three CMIP6 models indicate a predictive skill comparable to CMIP5 for the multimodel mean in general, and a distinct positive influence of the improved soil–hydrology scheme implemented in the MPI-ESM for CMIP6 in particular.

KEYWORDS: Water masses/storage; Soil moisture; Satellite observations; Forecast verification/skill

1. Introduction

Forecasting global or regional climatic conditions for several years into the future is now within reach after numerous scientific breakthroughs in the field of decadal climate prediction (Doblas-Reyes et al. 2013; Boer et al. 2016; Marotzke et al. 2016). Decadal prediction services operated by meteorological agencies provide unconditional forecasts for up to 10 years in advance by initializing Earth system models (ESMs) with observations or reanalysis data (Meehl et al. 2009). This initialization sets decadal predictions apart from long-term climate projections covering a century or more that are governed by boundary conditions such as the greenhouse gas concentrations or the solar activity. Projections therefore reproduce the climate variability in a statistical sense only, but offer no information about the actual conditions in the next 2–10 years ahead.

Due to its relevance for agricultural and water management decisions the information about a future evolution of surface temperatures and precipitation rates is very interesting for stakeholders. Therefore, the skill of present-day decadal prediction systems has been extensively assessed for state variables like sea surface and land temperatures (Corti et al. 2012; Bunzel et al. 2018), and also associated indices as the Atlantic multidecadal or Pacific decadal oscillations (Kim et al. 2012). Forecasting hydrometeorological quantities appears to be more challenging, with still limited forecast skill for precipitation (Mehrotra et al. 2014) and soil water availability (Yuan and Zhu 2018; Zhu et al. 2019). This is certainly related to the difficulties of accurately modeling those spatially and temporally highly variable quantities, but also to the limited availability of satellite and in situ observations that can be utilized for both model validation and calibration.

A satellite mission designed to map Earth's gravity field has been providing time variations in regional terrestrial water storage (TWS), which can be regarded as the integration of precipitation, evapotranspiration, and lateral runoff over time as described by the water balance equation. The Gravity Recovery and Climate Experiment (GRACE, in orbit from April 2002 to October 2017; Tapley et al. 2019) consists of two small twin satellites orbiting Earth at a very low altitude (less than 500 km) with a typical distance of about 220 km. Both satellites continuously measure the changes in their relative distance that are caused by spatial variations in Earth's gravitational attraction. Differences in those measurements between

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-20-0042.s1.

Corresponding author: Laura Jensen, laura.jensen@hcu-hamburg.de

DOI: 10.1175/JCLI-D-20-0042.1

subsequent overpasses are traced back to changes in water mass stored at or beyond Earth's surface. The observations from GRACE are being continued by the GRACE-FO mission (Flechtner et al. 2016; Kornfeld et al. 2019) launched in May 2018.

Data from GRACE were frequently used for the validation of both hydrological (Döll et al. 2014; Eicker et al. 2014; Güntner 2008; Syed et al. 2008) and land surface models (Scanlon et al. 2018; Zhang et al. 2017). The record has been compared also against long-term projections of ESMs (Rodell et al. 2018; Jensen et al. 2019), but rarely been used to evaluate decadal predictions. In an early attempt, Zhang et al. (2016) utilized GRACE-derived TWS to assess the effects of different initialization techniques on the quality of MPI-ESM hindcasts. In the present study, a GRACE-based TWS dataset is for the first time employed to evaluate a multimodel ensemble of decadal climate prediction experiments published in the context of phases 5 and 6 of the Climate Model Intercomparison Project [CMIP5 (Taylor et al. 2012) and CMIP6 (Eyring et al. 2016)]. The skill of the decadal hindcasts is assessed both globally and regionally by means of anomaly correlation and root-mean-square deviation (RMSD). We can demonstrate that the new observation type ''terrestrial water storage'' as available from the GRACE and GRACE-FO missions is suitable as additional dataset in the validation and/or calibration of climate model experiments. Since data from CMIP and GRACE are jointly available in only 9 years (2002–2011), we make use of a GRACE-like reconstruction of TWS, which expands the analysis time frame to 41 years.

2. Data and methods

a. *GRACE, GRACE-FO, and GRACE-REC*

From the sensor data collected by GRACE and GRACE-FO, it is possible to unambiguously quantify surface mass changes. By subtracting high-frequency mass variations (atmosphere and ocean non-tidal mass variability, tides in atmosphere, oceans, and solid Earth) and non-water-related processes (glacial isostatic adjustment, tectonic displacements), the water changes on land are isolated from this integrated signal. GRACE-derived TWS changes typically have a temporal resolution of one month and a spatial resolution of a few hundred kilometers. It is inherent to the measurement principle that GRACE-derived TWS changes contain all storage compartments (i.e., soil moisture, groundwater, snow, permafrost, glaciers, ice sheets, rivers, and lakes), and with GRACE alone they cannot be disaggregated into their different origins. GRACE observations are directly traced back to the measurement of time differences and are therefore not affected by long-term drifts and biases (Kim and Tapley 2002). Thus, satellite gravimetry can be regarded as a long-term stable observation technique for land water storage changes.

The currently available time series from GRACE and GRACE-FO range from April 2002 to November 2019. The majority of decadal hindcast experiments of CMIP5 are initialized only until 2010 (i.e., forecast year 1 equals 2011). Thus, only up to 2011 we can access model data for all forecast years (1 to 10). As this is crucial for our analysis, the effective overlap time span of GRACE/GRACE-FO with CMIP5 decadal hindcasts is just 9 years. Deriving forecast skill from only nine data points is likely dominated by random noise and robust results can hardly be expected. For example, a correlation coefficient of two time series with nine data points each would have to be larger than 0.67 to be significantly different from zero (with a significance level of 95%). To increase the overlap time span between observations and decadal hindcasts we make use of a century-long reconstruction of climate-driven water storage changes that is based on GRACE observations (GRACE-REC; Humphrey and Gudmundsson 2019).

By assuming that short-term anomalies of TWS are mainly driven by fluctuations in the relevant atmospheric drivers, Humphrey and Gudmundsson (2019) use precipitation and temperature data from atmospheric reanalyses to reconstruct past anomalies of TWS. The statistical model is based on the assumption that precipitation events have an exponentially decaying influence on the subsequent water storage that is governed by the temperature-dependent residence time of the water in the soil. Three parameters of the statistical model are calibrated for each grid cell against GRACE observations by means of a least squares adjustment: one parameter for the scale and two related to the residence time.

For this study, we use the reconstruction calculated with the Goddard Space Flight Center (GSFC) GRACE solution (Luthcke et al. 2013) and Global Soil Wetness Project phase 3 (GSWP3) precipitation and temperature (Kim 2017). As demonstrated by Humphrey and Gudmundsson (2019), GRACE-REC is close to the original GRACE observations within the overlapping period with a correlation of monthly global land averages larger than 0.75. In the yearly averaged time series that we use in our study the correlation is even higher with 0.92 (see online supplemental material section S1). GRACE-REC fits better to GRACE than TWS estimates from hydrological or land surface models in terms of correlation and Nash–Sutcliffe efficiency. Furthermore, GRACE-REC was evaluated against several observational datasets, including basin-scale water balances from ERA-Interim and runoff observations, as well as streamflow measurements. Particularly the comparison to streamflow measurements from 1901–2010 showed that even though GRACE-REC was calibrated to GRACE within the GRACE time span only, the correlation does not degrade for the earlier time spans, where no calibration data are available. Thus, we assume GRACE-REC to be a reliable estimate for water storage changes also for the years prior to the GRACE era.

The reconstruction is affected by several sources of uncertainty, including measurement and processing uncertainties in GRACE, structural model errors, and uncertainties in the precipitation and temperature data. To consider these spatially and temporally correlated errors, Humphrey and Gudmundsson (2019) derived in total 100 ensemble members of the GRACE-REC dataset by employing a spatial autoregressive noise model generating random realizations of the error structure. Thus, it is possible to derive realistic aggregated errors for basin-averaged time series such as the global land average. Although GRACE-REC is only a proxy for real GRACE observations, we consider it as a feasible replacement to demonstrate the value of a long TWS record for decadal prediction analysis: Not only is the

TABLE 1. Models used in the analysis. The upper five models take part in CMIP5; the lower three models take part in CMIP6. The name, institution (with country), reference, original spatial resolution, and the number of ensemble members for the decadal (Init) and the uninitialized (Hist) experiments are provided.

Name	Institution	Reference	Resolution	Init	Hist
		CMIP5			
Fourth Generation Canadian Coupled Global Climate Model (CanCM4)	Canadian Centre for Climate Modeling and Analysis (Canada)	von Salzen et al. (2013)	2.8°	10	5
NOAA's Geophysical Fluid Dynamics Laboratory Coupled Model, version 2.1 (GFDL-CM2p1)	NOAA Geophysical Fluid Dynamics Laboratory (United States)	Delworth et al. (2006)	2.5° × 2°	10	10
Hadley Centre Coupled Model, version 3 (HadCM3)	Met Office Hadley Centre (United Kingdom)	Gordon et al. (2000) and Pope et al. (2000)	3.75° × 2.5°	10	10
Model for Interdisciplinary Research on Climate, version 5 (MIROC5)	University of Tokyo, National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology (Japan)	Watanabe et al. (2010)	1.4°	6	3
Max Planck Institute Earth System Model, low resolution (MPI-ESM-LR)	Max Planck Institute for Meteorology (Germany)	Giorgetta et al. (2013) and Müller et al. (2012)	1.9°	3	3
		CMIP6			
Fourth Generation Canadian Coupled Global Climate Model (CanESM5)	Canadian Centre for Climate Modeling and Analysis (Canada)	Swart et al. (2019)	2.8°	20	25
Model for Interdisciplinary Research on Climate, version 6 (MIROC6)	University of Tokyo, National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology (Japan)	Tatebe et al. (2019)	1.4°	10	3
Max Planck Institute Earth System Model, version 1.2, high resolution (MPI-ESM1–2-HR)	Max Planck Institute for Meteorology (Germany)	Müller et al. (2018)	0.9°	5	2

correlation of GRACE-REC and the original GRACE observations for the yearly global land average very high (0.92), but also the evolution agrees very well with the GRACE time series lying within the error bounds of the reconstruction (see supplemental material section S1).

We note that GRACE-REC is derived from precipitation and temperature data and is thus not entirely observation-based despite of being calibrated against satellite gravity data. However, as none of the ESMs evaluated in this study is initialized or forced with GSWP3 data, we assume that GRACE-REC is a largely independent dataset for the comparison with water storage–related variables simulated by decadal hindcasts of coupled ESMs. The observational record of GRACE is being continued by GRACE-FO, and next-generation gravity missions are currently being prepared in the United States, China, and Europe (Pail et al. 2015) so that it is safe to assume that gravity observations will be available for validation, calibration (and possibly even initialization) of ESM decadal prediction experiments also in the future. We emphasize that GRACE-REC is used in this study only to extend the sample size for arriving at statistically more robust results. For the evaluation of individual forecasts from different operational decadal prediction systems, we always recommend using real satellite data from GRACE and GRACE-FO as readily available from, for example, the GravIS portal maintained by GFZ Potsdam (gravis.gfz-potsdam.de).

b. *CMIP5 decadal hindcasts*

CMIP5 models do not provide a standard output variable for terrestrial water storage. We therefore sum up the variables total soil moisture content (mrso) and surface snow amount (snw) to approximate modeled TWS (abbreviated as mTWS in the following). For the variables mrso and snw, five CMIP5 models provide monthly mean output of decadal hindcasts which were initialized every year with ocean temperature and salinity fields (Table 1). Four of these models are initialized from 1960 to 2010, while for one model the last initialization year is 2009. Each model experiment consists of 3–10 ensemble members usually generated using 1-day lagged fields for initialization. For further analysis we compute the ensemble mean per model as well as a multimodel mean (MMM) from all model ensemble means (39 members in total). For the MMM we also compute the spread as the weighted standard deviation from all ensemble members, propagating the uncertainty of the individual models to the uncertainty of the MMM giving each model equal weight (see supplemental material section S2).

Each hindcast runs for 10 years after its year of initialization. The mTWS anomalies for the months of the first full year after initialization (i.e., forecast year 1) are expected to be close to the observations because the influence of the initialization is still large. The mTWS anomalies for the second year after initialization (i.e., forecast year 2) are expected to fit a bit less to the observations than those from forecast year 1, but with a skillful forecasting system they should still fit better than a trivial forecast. With increasing lead time after initialization (forecast years 3–10) the forecast skill is expected to degrade further with respect to the observations. To assess the forecast time span up to which the decadal experiments still exhibit skill with respect to an uninitialized forecast, we build so-called forecast year time series. This means that we rearrange the mTWS anomalies from all decadal simulations with respect to their forecast year: Since the decadal simulations are initialized every year between 1960 and 2009 (at least), there exist forecast year 1 mTWS anomalies for all models for each year between 1961 and 2010, and if we keep only these first-year mTWS anomalies from each decadal hindcast we obtain a discrete time series consisting only of forecast year 1 mTWS anomalies. Analogously, forecast year 2 mTWS anomalies exist for each year between 1962 and 2011, constituting a forecast year 2 time series. This can be done for the other forecast years 3–10 as well. The first year for which the tenth forecast year exists, is 1970 (10 years after the first initialization in 1960). Thus, the common time span where each forecast year between 1 and 10 is available is 1970 to 2010, hence this is the time span for which we perform our further analysis. The forecast year time series derived from decadal hindcasts are referred to as initialized simulations (Init) in this study.

As a reference for the skill assessment we use mTWS time series from 1970 to 2010 obtained from historical runs of the same CMIP5 models. Historical CMIP5 experiments are typically initialized from an arbitrary point of a quasi-equilibrium control simulation. Their starting date is set to 1850, and simulations are forced by observations of, for example, solar insolation, greenhouse gas emissions, and land cover change (Taylor et al. 2012). The historical experiments in CMIP5 usually end in 2005, hence for our analysis we extend them until 2010 with data from CMIP5 projections under the representative concentration pathway scenario 4.5 (RCP4.5, i.e., assuming a moderate increase in greenhouse gas concentration and radiative forcing until 2100). As the conditions in 1850 have virtually no influence on the simulated data for the years 1970–2010 we refer to these concatenated reference runs as uninitialized or historical simulations (Hist) in the following. Please note that for CanCM4, snw is not stored in the CMIP5 archive for both historical and RCP4.5 simulations, so that we use the corresponding runs from CanESM2 instead, which consists of CanCM4 coupled to a terrestrial and ocean carbon model. Also for Hist we compute ensemble means per model and a multimodel mean from 31 members in total. All monthly model output grids are remapped to a common 2° × 2° geographical grid.

c. *Calculation of anomalies*

To be independent of seasonal variations and to exclude biases due to the time of initialization of the decadal experiments, the monthly time series for GRACE-REC, Init, and Hist are averaged to annual sampling. Subsequently, from each time series the linear trend and bias for the time span 1970–2010 are removed to obtain anomalies. We restrict our analysis to those detrended values since the linear drifts present in the GRACE-REC time series originate solely from trends in the precipitation dataset used for the reconstruction. Thus, they are not fully representative for all long-term changes in TWS, since long-term changes in runoff and evapotranspiration are not considered. Furthermore, the trends are different for different versions of the GRACE-REC dataset that use different reanalyses, and are not everywhere similar to the trends in the original GRACE observations, which also capture changes in deep groundwater. In addition, there is still a large intermodel spread regarding soil moisture and snow trends in CMIP5 models, which restricts consensus between trends in GRACE-based TWS and mTWS from CMIP5 to selected regions only (Jensen et al. 2019).

We recall that TWS and mTWS anomalies that remain after removing the linear trend do not entirely represent the same physical entity. Model-based mTWS does not include surface water variability in rivers and lakes, which are typically represented by a river routing module in ESMs but are not stored in the CMIP5 archive. Furthermore, mTWS does not capture anthropogenic interventions on the water cycle such as groundwater abstraction or dam building, which is an emerging signal in the GRACE TWS observations (Voss et al. 2013). In addition to the incomplete representation of TWS in ESMs, the GRACE-REC dataset might be biased in some regions by non-water-related processes, such as glacial isostatic adjustment (GIA) and tectonic deformations. To account for such conceptual differences in TWS and mTWS we exclude in our analysis regions that are strongly affected by surface water variability, groundwater abstraction, and earthquakes (about 7% of the land surface without Greenland and Antarctica; see supplemental material section S3). GIA causes a long-term linear mass trend, hence not influencing the annual anomalies. The soil depth realized in ESMs is typically limited to a constant depth of a few meters, which probably is not representative for the full water holding capacity everywhere. However, a certain fraction of deeper soil layers, groundwater, and surface water can be implicitly contained in total soil moisture content as the water budget is approximately closed in the CMIP5 models (Liepert and Lo 2013) and water transport to ocean and atmosphere is limited. But groundwater dynamics beneath the soil layer and groundwater–soil interactions are not represented in the models and thus their feedback on the climate system is not considered in the CMIP5 ESMs, possibly leading to systematic deficits. Even though mTWS might not capture the full magnitude of the water storage variability at least the relative changes in the anomalies should be similar, as a drying or wetting of the upper soil layers is often reflected in a general drying or wetting of all water storage compartments (Swenson et al. 2008). Thus, at least in terms of Pearson's correlation coefficient that is used in the following sections as one of our evaluation metrics, the different magnitudes of TWS and mTWS should be of minor consequences for the results.

3. Results

a. *Global average*

To assess the general skill of CMIP5 decadal hindcasts regarding mTWS we first analyze time series of the global land

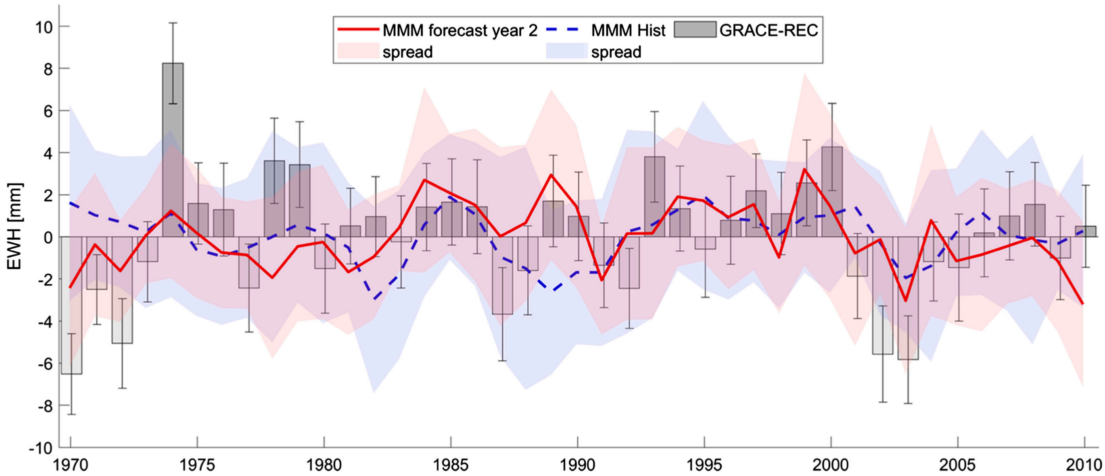


FIG. 1. Global land average (Greenland, Antarctica, and regions highly affected by surface waters, groundwater abstraction, and earthquakes excluded) annual time series for 1970–2010 for GRACE-REC TWS anomalies (gray bars), forecast year 2 initialized hindcast mTWS of multimodel mean (MMM; red line), and mTWS MMM of uninitialized simulations (dashed blue line). Corresponding error bounds computed as the (weighted) standard deviation of the respective ensemble members are indicated by thin black error bars (GRACE-REC) and light red and light blue shaded areas (MMM Init and Hist, respectively).

average (excluding Greenland and Antarctica, and regions highly affected by surface waters, groundwater abstraction, and earthquakes as defined in the supplemental material) calculated from the yearly mTWS anomalies in the time span 1970–2010. For illustration, in Fig. 1 the global mean GRACE-REC anomaly time series expressed in equivalent water height (EWH) is displayed with gray bars and corresponding error bars obtained from 100 ensemble members. It is overlaid by the global mean forecast year 2 anomaly time series of the multimodel mean (MMM; red line) and the global mean anomaly time series of the MMM from the uninitialized runs (dashed blue line). The respective model spreads are depicted in light red and light blue shading. The mean spread of the forecast year 2 Init time series is 3.44 mm EWH, and for the Hist time series it is 3.77 mm EWH. Both values exceed the mean spread of the GRACE-REC time series (2.05 mm EWH), and thus we consider GRACE-REC a reliable reference for evaluation of model results. Furthermore, the Init spread is smaller than the Hist spread, which hints at a superior reliability of Init predictions over Hist experiments for forecast year 2. We note that the root-mean-square (RMS) of the global mean time series of the Init run (1.56 mm EWH) is only about half as large as for GRACE-REC (2.92 mm EWH), and for the Hist run (1.22 mm EWH) even smaller. This might point toward some skill in representing variability of the initialized predictions compared to uninitialized runs. One reason for smaller variability in the models (compared to GRACE-REC) might be the incomplete representation of TWS in CMIP5 models discussed above. Another reason is the tendency of multimodel means to smooth out temporal anomalies via ensemble averaging, which is a known issue in seasonal and decadal modeling (Smith et al. 2019). Several approaches for rescaling forecast anomalies have been proposed; however, the discussion about the best method is still ongoing, so none of those methods is implemented here.

The correlation (which is unaffected by the magnitude of the signal) of the GRACE-REC time series with the MMM forecast year 2 time series is 46%, which is substantially higher than the correlation with the MMM Hist time series (15%). Furthermore, the RMSD between the observational time series and the forecast year 2 initialized time series is smaller than for the Hist time series (2.58 vs 2.95 mm EWH). We repeat the computation of the correlation and RMSD between the GRACE-REC anomaly time series and the model time series for all forecast lead times from 1 to 10 years (Fig. 2). In addition to the MMM (black lines) we also compute the correlations and RMSD for the ensemble means of the five individual models (colored lines in Fig. 2). As expected, the correlation generally decreases with increasing forecast year. For the MMM the initialized hindcasts exceed the uninitialized runs (stippled lines in Fig. 2) in terms of correlation for the first three forecast years (0.64, 0.46, and 0.24 vs 0.15). From forecast year 4 onward no clear improvement of Init over Hist is found. The same holds for the RMSD (Fig. 2b), which is clearly smaller for Init than for Hist for the first two forecast years (2.23 and 2.58 mm vs 2.95 mm) and very slightly smaller for the third forecast year (2.93 mm).

For the individual models (colored lines) the correlation–forecast year relationship is noisier, but for the majority also at least the first three forecast year correlations are above the Hist correlation of the respective model. Exceptions are the HadCM3 and the MPI-ESM-LR: for these models the Hist correlation is already comparably high (0.29 and 0.31), and only the first (MPI-ESM-LR) and respectively second (HadCM3) forecast years are above this value. In the HadCM3 some later forecast years are also above the Hist correlation, but this is probably not a robust result. For forecast year 1 and 2 the correlation for the MMM is higher than all individual model correlations (black line above colored lines), and for the RMSD the MMM has the lowest values compared to the individual models. This suggests that using an ensemble of different models for forecasting mTWS is preferable over using

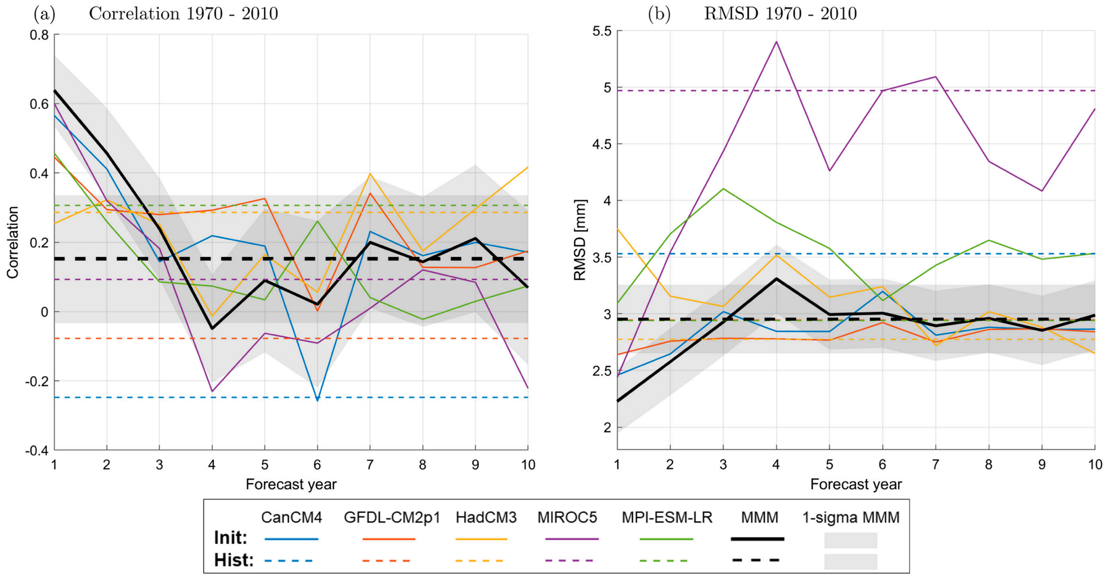


FIG. 2. (a) Correlations (as a function of forecast lead time) of the global mean GRACE-REC TWS anomaly time series with the global mean forecast year mTWS time series from decadal hindcasts (solid lines) and the uninitialized time series (dashed lines) for the time span 1970–2010. Colored lines indicate individual ESMs; the solid black line denotes the multimodel mean (MMM) of the five ESMs. Greenland and Antarctica are excluded. (b) As in (a), but for RMSD. Light gray shaded areas denote the standard deviations of the MMM correlations and RMSDs.

one specific model alone. A positive relationship between the size of the ensemble and the correlation of the ensemble mean with the observations was already found in decadal hindcasts for other variables (e.g., temperature and precipitation; Smith et al. 2019) and here we show that this also applies for TWS. The time scale of 3 years that we identify for the improved prediction skill of the Init over the Hist simulations is largely consistent with a study from Yuan and Zhu (2018), who analyzed the maximum lead times where initial conditions prevail over meteorological forcings in TWS predictability and found it to be shorter or equal to 3 years in 79% of the land area (Greenland, Antarctica, and desert regions excluded), and shorter or equal to 5 years in even 89%.

For the correlation and RMSD values of the MMM we perform an error propagation considering the spread of the ensemble members of GRACE-REC and of the Init/Hist runs (see supplemental material section S2). The resulting error bounds are displayed in light gray in Fig. 2 representing the standard deviation of the correlation and RMSD values (1-sigma). As expected from the large model spread, the uncertainties of the correlations and RMSDs for the global average are quite large and only in the first forecast year a clear separation between Init and Hist simulations is seen. Thus, although there is some indication that forecast year 2 and 3 exhibit forecast skill (the correlations are higher than those for forecast year 4–10, and higher than the Hist correlations; RMSD values are respectively lower), at this time no clear conclusion can be drawn about the robustness of this result. The relatively large error bounds also arise from a limited number of data points (41) from which correlation and RMSD are calculated and thus will decrease with an increasing number of hindcast experiments.

For the global average the Init predictions outperform the Hist experiments for the first two to three forecast years. To quantify the added value of especially the second and third forecast years of the decadal predictions we compare the results to those from a persistent forecast (Fig. 3). This means that instead of using the actually predicted TWS state we retain the TWS state of the first forecast year also for the second, third, and so on up to tenth forecast year. Keeping the prediction for the first forecast year for the next couple of years would—in case of having a similar quality as the decadal predictions—be a cheap alternative for dynamic forecasting of TWS from an ESM integration. However, when calculating the correlation of the global average GRACE-REC time series with the MMM persistent forecast, it shows that for forecast years 2 and 3 it is substantially lower than for the decadal predictions, whereas the RMSD is higher (Fig. 3). This further supports our earlier conclusion that the decadal predictions have an actual forecast skill for mTWS beyond the first forecast year. The light gray and light red bounds around the curves in Fig. 3 denote the 1-sigma error boundary of the correlation coefficients and the RMSDs, calculated via variance propagation of the ensemble spread of GRACE-REC and the ESMs (light gray same as in Fig. 2). Due to the large overlap of the error bounds especially in the third year these results still remain somewhat arguable. In addition to the rather short time span that contributes to the uncertainty, it is mainly caused by the spread of the ESM results.

To test if the model spread is an appropriate measure for the prediction uncertainty (Goddard et al. 2013) we calculate the temporal mean of the spread for the Init and Hist runs and compare it to the standard deviation of the differences between

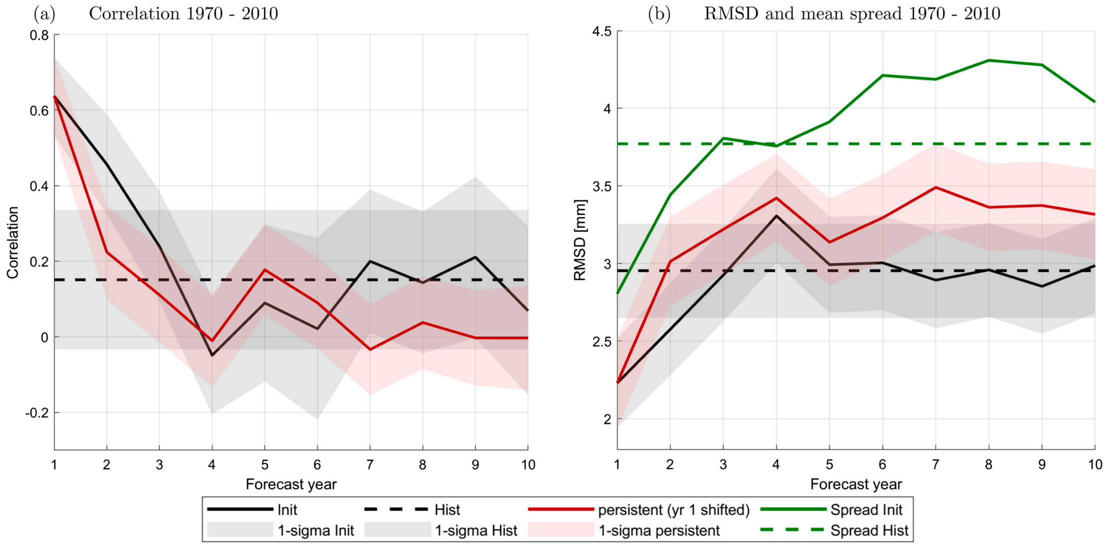


FIG. 3. Comparison of mTWS Init simulations (solid black, as in Fig. 2) and a persistent forecast (red) for the multimodel mean (MMM) global average time series for the time span 1970–2010 in terms of (a) correlation and (b) RMSD. The green lines in (b) denote the mean spread of the MMM for Init and Hist. The light gray (as in Fig. 2) and red bounds indicate the 1-sigma error boundaries.

MMM and observation anomalies (i.e., the RMSD). In a perfectly calibrated prediction system the two measures should be the same (Palmer et al. 2006). However, here the model spread overestimates the RMSD (cf. green lines to black lines in Fig. 3b) by a factor of 1.1 to 1.5 (mean 1.3). This indicates that the inhomogeneity between the different CMIP5 models is still too large for reliable forecasts of mTWS, which was similarly found for instance by Goddard et al. (2013) and Doblas-Reyes et al. (2013) for other variables (temperature, precipitation). As a result also the error boundaries of the correlation and RMSD values probably are rather pessimistic estimates. We believe that five models with a total number of 39 ensemble members do not represent a perfectly calibrated system, and a one-to-one match thus cannot be expected. An increased number of ensemble members and further model developments might improve the reliability (see section 3c). Apart from the (rather constant) factor between Init model spread and Init RMSD, the evolution of the two measures over the different forecast years is increasing in parallel, which means that the model spread is generally reflecting the influence of the initialization on the forecast quality. Furthermore, the Init spread is smaller than the Hist spread for the first two forecast years, consistent with the findings for correlation and RMSD and further strengthening the conclusion of a global mean forecast skill of decadal mTWS hindcasts for the first two to three forecast years.

b. *Regional analysis*

In addition to the analysis of the global mean, also regional skill assessments are performed. For Fig. 4 we calculate annual time series averaged over different Köppen–Geiger climate zones (Peel et al. 2007). In equatorial regions (22% of land area) the initialized runs clearly outperform the uninitialized runs for the first three forecast years (Fig. 4a). For these years the MMM correlation is substantially higher than the global mean correlation (0.90, 0.64, and 0.38 vs 0.64, 0.46, and 0.24; cf. Fig. 2a) and also exhibits substantially smaller error bounds. The good forecast skill in equatorial regions is caused by a generally deeper soil depth compared to the other climate zones and correspondingly a longer soil moisture memory of the initialization (Stacke and Hagemann 2016). In the other climate zones only the first forecast year shows a clear predominance of Init over Hist runs, thus the forecast skill for TWS seems to be limited to shorter lead times in these regions (Figs. 4b–d). In temperate regions (16% of land area) the first year's correlation is slightly higher than for the global mean correlation (0.74 vs 0.64), whereas for arid and polar regions it is lower (0.32 and 0.40). The reason for the poor performance in arid regions (36% of land area) could be related to the generally limited presence of water combined with sporadic rain events. In polar regions (26% of land area) the low correlations might be due to a limited or even missing representation of frozen soil and surface water in ESMs and the generally more complex hydrological processes related to snow accumulation and melting. Temperate regions only cover a small percentage of the land area, so the aggregation area might be too small to yield a reliable result.

For a more detailed regional analysis of forecast skill we compute global maps of correlation for the GRACE-REC with the Init and Hist MMM time series (Fig. 5). From the visual comparison of the Hist correlation map (Fig. 5a) with the MMM forecast year 1 correlation map (Fig. 5b) we conclude a general success of the initialization, as its correlation is much higher than without initialization. This shows that initialization has a direct positive effect not only on the respective initialized variables (e.g., ocean temperature and salinity) but also on derived variables such as mrso and snw. The MMM forecast

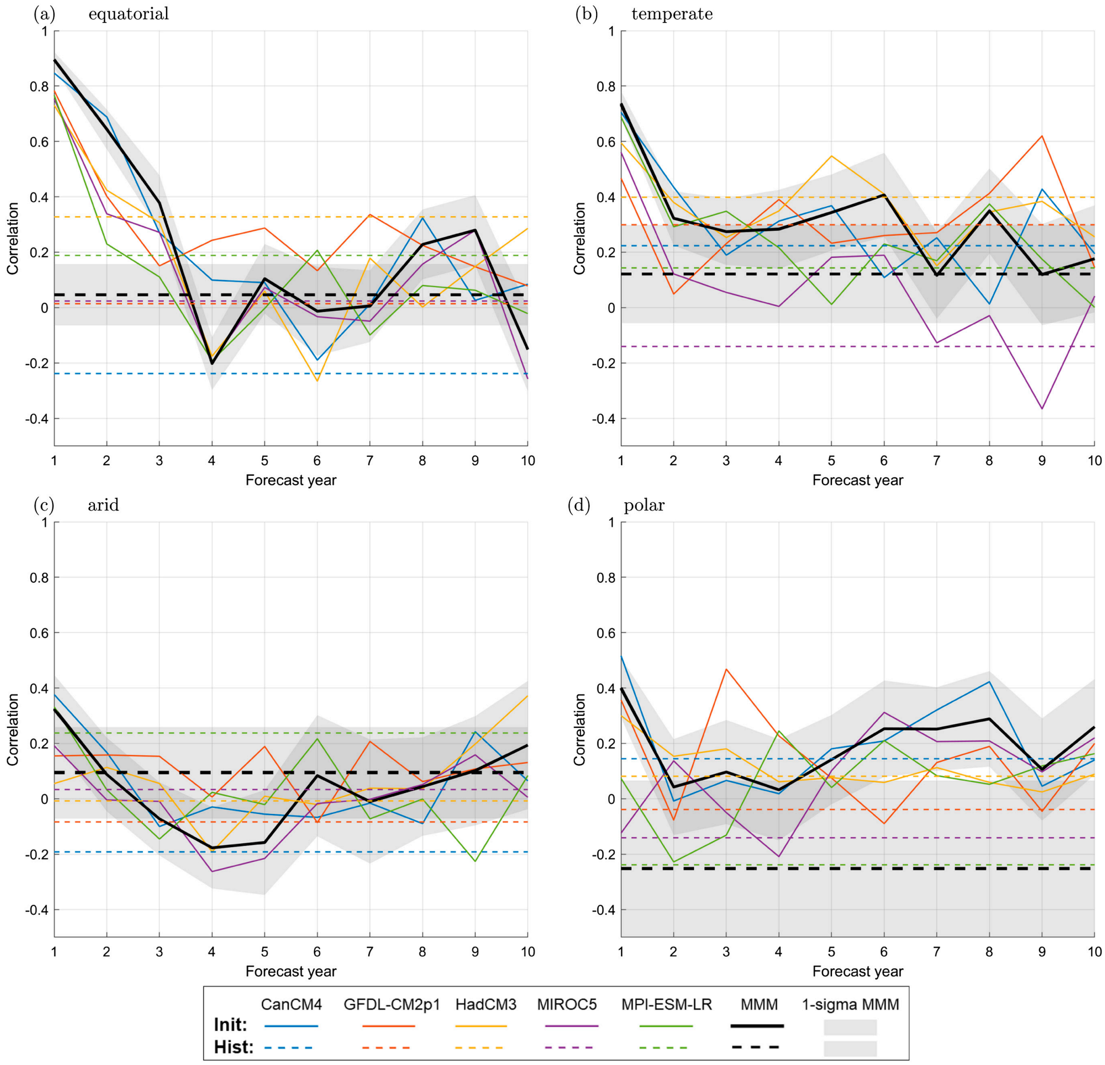


FIG. 4. Correlations (as a function of forecast lead time) of the GRACE-REC TWS anomaly time series with the forecast year mTWS time series from decadal hindcasts (solid lines) and the uninitialized time series (dashed lines) averaged over different climate zones for the time span 1970–2010. Colored lines indicate individual ESMs; the solid black line denotes the multimodel mean (MMM) of the five ESMs. Light gray shaded areas denote the standard deviations of the MMM correlations.

year 2 correlation map (Fig. 5c) also exhibits a higher fraction of positive correlations than the Hist correlation map.

To more objectively analyze the maps, we calculate for each map (and for the maps for forecast year 3–10; not shown) the percentage of global land area exhibiting a significantly positive or negative correlation (Fig. 5d, blue curves). Furthermore, we obtain the percentage of significantly positive or negative correlation within the equatorial climate zone as defined in the Köppen–Geiger classification scheme (Fig. 5d, red curves). The significance of the correlation coefficients was tested for a confidence level of 95%. For forecast years 1 and 2 of the initialized hindcasts, the global land area fraction being significantly positive is clearly above the corresponding value from Hist (38% and 16% vs 9%). For forecast year 3, the fraction (12%) is still higher than for the Hist simulations and all longer lead times between 4 and 10 years (max. 10%). Yet, the difference from the later forecast years is not as distinctive as for the first two forecast years. Thus, a general grid-scale forecast skill of CMIP5 decadal predictions for TWS of 3 years (or even more) is not identified. However, when focusing on the equatorial climate zone only, the percentage of significantly positive correlations in forecast year 3 is clearly higher than for the later forecast years (15% vs a maximum of 11%) and also compared to Hist (10%). This confirms the results from Fig. 4a and suggests that the predictive skill in equatorial regions is higher than for other climate zones, possibly due to a longer-lasting influence of the initialization caused by an increased soil water memory time in these regions. The results for the significantly negative correlations (light blue and red curves in Fig. 5d) largely reflect the findings for the significantly positive correlations and thus are not further discussed here.

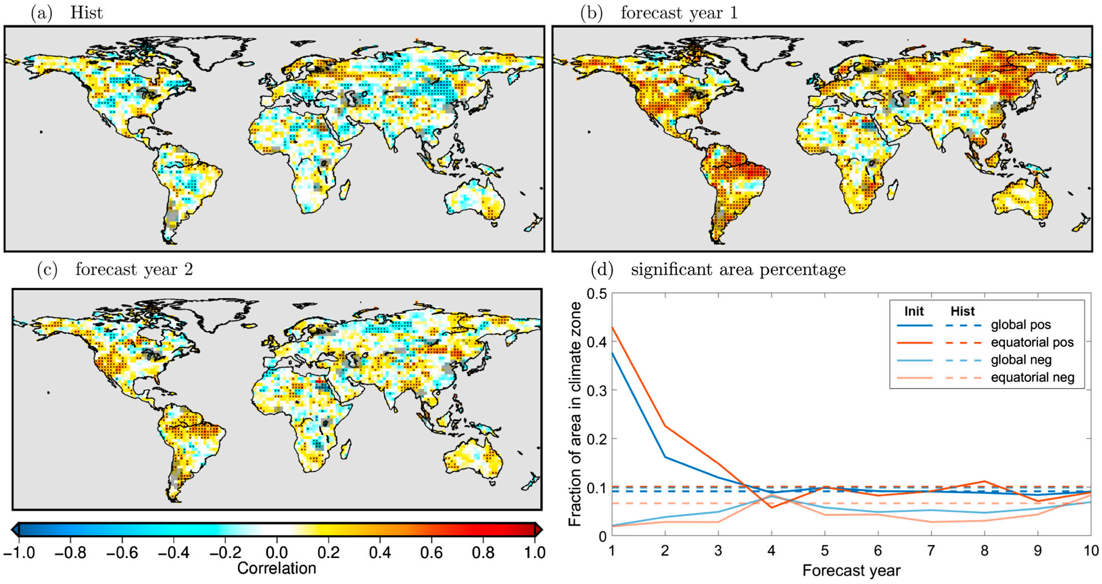


FIG. 5. Global maps of correlation of the GRACE-REC TWS anomaly time series with (a) uninitialized, (b) forecast year 1, and (c) forecast year 2 MMM mTWS anomaly time series for the time span 1970–2010. Stippled areas indicate significant correlation. Areas strongly affected by surface waters, groundwater abstraction, or earthquakes are shaded in gray. (d) Percentage of land area with significantly positive and negative correlation as function of forecast lead time for the global land area (blue) and the equatorial climate zone (red).

c. *A first look into CMIP6*

Phase 6 of the Coupled Model Intercomparison Project (CMIP6; Eyring et al. 2016) started a few years ago, and the results from the various efforts are currently being made available. Besides major changes in the organization of the simulations, the participating ESMs were subject to further developments of their physical and numerical schemes. One element of CMIP6 is the Decadal Climate Prediction Project (DCPP; Boer et al. 2016) which defines the experiment setup for initialized simulations. Compared to CMIP5, more frequent initialization dates and larger ensemble sizes are expected to increase the robustness of the predictions. However, the choice of methods to initialize the simulations and to generate ensembles is still left to the individual research groups and is not specified by DCPP.

At the time of writing, CMIP6 decadal hindcasts and corresponding Hist simulations are available for the variables mrso and snw from four ESMs. As the IPSL-CM6A-LR does not provide yearly-initialized decadal runs for its predecessor model from CMIP5, we restrict the analysis to three models (CanESM5, MIROC6, and MPI-ESM1–2-HR; see Table 1) with 35 ensemble members (30 members for the Hist simulations) in total. From these we compute ensemble means per model. We also calculate a multimodel mean from the ensemble means of the three models together with the weighted MMM spread. We apply the same processing as before: building forecast year time series, and calculating correlations and RMSDs from the global mean Init and Hist with the global mean GRACE-REC TWS time series depending on the forecast year. Subsequently, we compare the results from the CMIP6 hindcasts of the three different models to the CMIP5 hindcasts of the respective predecessor models (CanCM4, MIROC5, and MPI-ESM-LR). We also compare the MMM from the three CMIP6 models to the MMM of the three corresponding CMIP5 models (Fig. 6).

Interestingly, the forecast year 1 correlations for the CMIP6 hindcasts are smaller than those for the CMIP5 hindcasts for two of the three models and the MMM (see Fig. 6, left) and the RMSDs in forecast year 1 are larger in CMIP6 vs CMIP5 hindcasts (see Fig. 6, right). However, with just three models providing data at this time, it is not yet possible to trace this behavior to a common source such as changes in the initialization strategy (full-field vs anomaly), the addition of further variables for initialization, changes in the initialization date, or simply the model resolution. The forecast year 2 correlations, however, are larger for CMIP6 than for CMIP5 for all three models and the MMM; the RMSD is smaller only for CanESM5 and MPI-ESM1–2-HR. For forecast year 3, the results again vary from model to model: CanESM5 and MIROC6 degrade relative to CanCM4 and MIROC5; and MPI-ESM1–2-HR improves substantially over its predecessor. The deviations between the three models result in slightly degraded forecast year 3 correlations and RMSDs for the MMM when progressing from CMIP5 to CMIP6. Concluding from only three models so far, the forecast skill of decadal mTWS predictions in CMIP6 for the first three forecast years seems to be on a similar level to that in CMIP5. However, the differences between the model generations depend on the respective model: for MPI-ESM (Figs. 6e,f) substantial improvements are documented between CMIP5 and CMIP6, but not for the other two models.

Generally, the CMIP5 MMM correlation curve (Fig. 6g) exhibits a clear linear decay of the correlation from forecast year 1 to 3 approaching the level of the Hist correlation for the

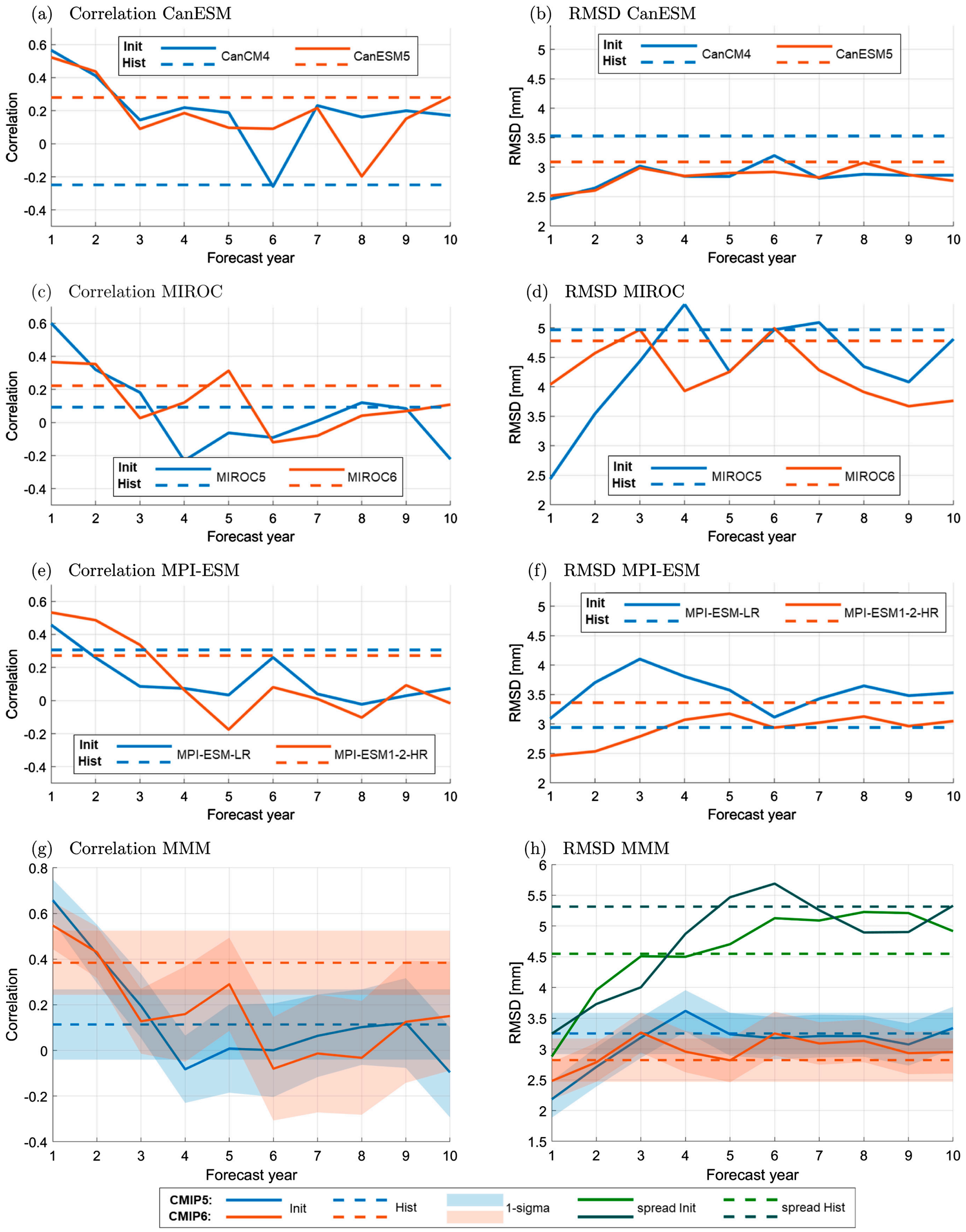


FIG. 6. Comparison of CMIP5 (blue) and CMIP6 (red) mTWS decadal hindcasts. (a),(c),(e) Correlation of the global mean GRACE-REC TWS time series and the Init and Hist mTWS time series as function of forecast time for three different ESMs for the time span 1970–2010. (g) The MMM from the three models above together with the 1-sigma error boundaries. (b),(d),(f),(h) As in (a), (c), (e), and (g), but for RMSD. Green lines in (g) indicate the mean spread of the ensemble members around the MMM.

later forecast years. This led us to the conclusion of a mTWS forecast skill limited to about 2–3 years (section 3a). However, for CMIP6 the shape of the correlation curve is not that distinctive: after a drop from forecast year 1 to 3 the correlations rise again for forecast years 4 and 5 before dropping to about the level of the CMIP5 correlations. This might be an indication for possible predictability beyond forecast year 3 in CMIP6 decadal predictions, but as only three models are evaluated here, the result is certainly not very robust. The MMM Hist correlation of the CMIP6 simulations is substantially higher than for CMIP5. Thus, in CMIP6, the Init simulations are already inferior to the Hist simulations after the second forecast year. Furthermore, for the later forecast years the Init correlations do not—as in CMIP5—approach the level of the Hist correlation but rather drop down to the CMIP5 correlation level. When several other modeling groups have finally published their decadal hindcast simulations, the results for the MMM and thus the conclusions on decadal prediction skills of CMIP6 regarding mTWS in general need to be confirmed.

Regarding the error bounds of the MMM correlation and RMSD curves, we note that the error boundaries do not notably decrease from CMIP5 to CMIP6. Furthermore, the mean overestimation of the RMSD by the model spread (green lines in Fig. 6h) also rises from a factor of 1.49 for CMIP5 to a factor of 1.60 for CMIP6. The reason might be that the increased model complexity involved in CMIP6 is also reflected in larger disagreements between model results and larger mean ensemble spreads. An indication for an improved forecast reliability in forecast years 2 and 3 in CMIP6 is the smaller spread of CMIP6 compared to CMIP5 in these years, leading to a convergence of model spread and RMSD. But the number of ensemble members is probably still too small to act on the assumption of a well-calibrated prediction system and to conclusively judge the prediction quality.

In Fig. 6 it is striking that, in contrast to CanESM and MIROC (Figs. 6a–d), the MPI-ESM metrics for forecast years 1–3 are much improved from CMIP5 to CMIP6 (Figs. 6e,f). The reason might be that for CMIP6 in the MPI-ESM a new five-layer soil–hydrology scheme was implemented that allowed for the separation of the soil into a top layer, root zone, and deep soil layer with physically distinct processes (Hagemann and Stacke 2015) while only a simple one-layer scheme was employed in the CMIP5 version of the MPI-ESM. This modification was already shown to improve surface temperatures (Bunzel et al. 2018) and affect soil moisture memory (Stacke and Hagemann 2016). Furthermore, the CMIP6 hindcasts are integrated with higher horizontal resolution than the CMIP5 ones (0.9° vs 1.9°). In contrast, for CanESM5 and MIROC6 no substantial changes were made either in the land surface component or in the spatial resolution compared to their predecessors from CMIP5. This might be an indication that incorporating more soil layers and a deeper soil depth in coupled ESMs has a positive impact on the prediction skill of decadal prediction regarding water storage–related variables.

4. Summary

We analyzed the forecast skill of decadal predictions from five yearly-initialized CMIP5 coupled Earth system models (Table 1) with respect to terrestrial water storage (TWS) related variables total soil moisture content (mrso) and surface snow amount (snw). We made use of a global reconstruction of climate-driven TWS changes (GRACE-REC; Humphrey and Gudmundsson 2019) that is based on observations from the satellite mission GRACE to carry out a skill assessment over 41 years in total (1970–2010). Skill was evaluated with respect to different yearly forecast horizons. Thus, we created forecast year time series from the yearly-initialized hindcasts (referred to as Init simulations) for the ensemble means of the individual models as well as for the multimodel mean (MMM) of the five models. As a reference we used the uninitialized (Hist) experiments (historical and RCP4.5 simulations) from the respective models. Afterward, we computed yearly-averaged anomaly time series (i.e., linear trend and bias removed) for the time span 1970–2010 from Init, Hist, and GRACE-REC.

The skill assessment was carried out on global and regional scales. We found that for the global land average of the MMM and the majority of the individual models the Init simulations outperform the Hist runs for the first three forecast years in terms of correlation and RMSD. We also deduced that the use of the MMM is preferable over individual models as the correlation is highest (RMSD is lowest) for the MMM in the first two forecast years and the general shape of the correlation curve is most distinct (monotonically decaying for the first 3 years and approximating the Hist level afterward) whereas the curves for the individual models are noisier. The maximum time of 3 years for the predominance of Init over Hist simulations is consistent with a study by Yuan and Zhu (2018), who found TWS predictability to be maximal 3 years for 79% of the land area. To demonstrate the actual forecast skill of the second and third forecast year we showed that the MMM global mean Init correlations for these years are also higher than those obtained from a persistent forecast, thereby underlining the added value of dynamic forecasts derived from ESM model runs with respect to trivial forecasts. We also analyzed if the ensemble spread around the MMM global mean is adequately representing the prediction uncertainty by comparing it to the RMSD between MMM and GRACE-REC anomalies. We found that the model spread generally reflects the rise of the RMSD with increasing forecast year, but overestimates it by a factor of about 1.3. This might be due to the relatively small ensemble size.

In the regional analysis we repeated the skill assessment for time series averaged over different climate zones. While in arid, temperate, and polar regions the results for the Init simulations are degraded in comparison to the global analysis, in the equatorial climate zone much higher correlations and smaller RMSDs were found. Even for forecast year 3, a clear prediction skill at 2° grid cell scales was documented in the equatorial climate zone. This is related to generally larger soil depths and thus longer soil memories in these regions leading to a longer-lasting influence of the initialization. From the 2° global maps, a general success of the initialization in forecast year 1 was identified (38% of land area exhibits significantly positive correlation, compared to 9% for the Hist runs). However, a general regional prediction skill for TWS for lead times longer than 2 years is not found in CMIP5.

We also assessed the forecast skill of decadal hindcasts already available for three CMIP6 models (Table 1) and their MMM, and compared the results from those of the respective CMIP5 models. The general level of prediction skill of the MMM global average for the first three forecast years was found to be similar for CMIP5 and CMIP6 from only three models available so far. An improved reliability of CMIP6 in the early forecast years might be indicated by the smaller mean ensemble spread compared to CMIP5. When looking at individual models, we noticed a clear improvement from CMIP5 to CMIP6 for MPI-ESM only, which might be due to the fact that in MPI-ESM a new five-layer soil–hydrology scheme was implemented for CMIP6, whereas for MIROC and CanESM no significant changes of the soil scheme were made. This indicates a positive impact of a multilayer hydrology scheme on the predictive skills of decadal simulations regarding TWS.

The current overlap time span between GRACE observations and CMIP5 decadal predictions is only 9 years, which is too short for a robust skill assessment. Hence, a global reconstruction of TWS extending back to 1970 has been used in this study to demonstrate the potential value of satellite gravity data for the assessment of decadal climate prediction. With more hindcast experiments from CMIP6 and a growing data record from GRACE-FO a direct comparison of satellite data with the results from ESM experiments at interannual to decadal scales will be possible very soon. Since satellite gravimetry senses mass anomalies independently of its surface exposure and physical condition, it is equally able to record changes in snow storage, soil moisture, and deep groundwater, thereby providing information about relative changes in the amount of available water at large spatial scales on the globe equally well in both tropical and polar climates.

Acknowledgments. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table 1 of this paper) for producing and making available their model output (e.g., at https://esgf-data.dkrz.de/search/cmip5-dkrz/). For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. Furthermore, we thank Vincent Humphrey and Lukas Gudmundsson for producing and making available the GRACE-REC data set.

Data availability statement. Raw CMIP5 and CMIP6 data are publicly available, e.g., on https://esgf-data.dkrz.de/search/cmip5-dkrz/ and https://esgf-data.dkrz.de/search/cmip6-dkrz/. The GRACE-REC data are stored on https://doi.org/10.6084/m9.figshare.7670849. Derived data supporting the findings of this study are available from the corresponding author upon request.

REFERENCES

Boer, G. J., and Coauthors, 2016: The Decadal Climate Prediction Project (DCPP) contribution to CMIP6. *Geosci. Model Dev.*, **9**, 3751–3777, https://doi.org/10.5194/gmd-9-3751-2016.

Bunzel, F., W. A. Müller, M. Dobrynin, K. Fröhlich, S. Hagemann, H. Pohlmann, T. Stacke, and J. Baehr, 2018: Improved seasonal prediction of European summer temperatures with new five-layer soil-hydrology scheme. *Geophys. Res. Lett.*, **45**, 346–353, https://doi.org/10.1002/2017GL076204.

Corti, S., A. Weisheimer, T. N. Palmer, F. J. Doblas-Reyes, and L. Magnusson, 2012: Reliability of decadal predictions. *Geophys. Res. Lett.*, **39**, L21712, https://doi.org/10.1029/2012GL053354.

Delworth, T. L., and Coauthors, 2006: GFDL's CM2 global coupled climate models. Part I: Formulation and simulation characteristics. *J. Climate*, **19**, 643–674, https://doi.org/10.1175/JCLI3629.1.

Doblas-Reyes, F. J., and Coauthors, 2013: Initialized near-term regional climate change prediction. *Nat. Commun.*, **4**, 1715, https://doi.org/10.1038/ncomms2704.

Döll, P., H. M. Schmied, C. Schuh, F. T. Portmann, and A. Eicker, 2014: Global-scale assessment of groundwater depletion and related groundwater abstractions: Combining hydrological modeling with information from well observations and GRACE satellites. *Water Resour. Res.*, **50**, 5698–5720, https://doi.org/10.1002/2014WR015595.

Eicker, A., M. Schumacher, J. Kusche, P. Döll, and H. M. Schmied, 2014: Calibration/data assimilation approach for integrating GRACE data into the WaterGAP Global Hydrology Model (WGHM) using an ensemble Kalman filter: First results. *Surv. Geophys.*, **35**, 1285–1309, https://doi.org/10.1007/s10712-014-9309-8.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016.

Flechtner, F., K.-H. Neumayer, C. Dahle, H. Dobslaw, E. Fagiolini, J.-C. Raimondo, and A. Güntner, 2016: What can be expected from the GRACE-FO laser ranging interferometer for Earth science applications? *Surv. Geophys.*, **37**, 453–470, https://doi.org/10.1007/s10712-015-9338-y.

Giorgetta, M. A., and Coauthors, 2013: Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5. *J. Adv. Model. Earth Syst.*, **5**, 572–597, https://doi.org/10.1002/jame.20038.

Goddard, L., and Coauthors, 2013: A verification framework for interannual-to-decadal predictions experiments. *Climate Dyn.*, **40**, 245–272, https://doi.org/10.1007/s00382-012-1481-2.

Gordon, C., C. Cooper, C. A. Senior, H. Banks, J. M. Gregory, T. C. Johns, J. F. B. Mitchell, and R. A. Wood, 2000: The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Climate Dyn.*, **16**, 147–168, https://doi.org/10.1007/s003820050010.

Güntner, A., 2008: Improvement of global hydrological models using GRACE data. *Surv. Geophys.*, **29**, 375–397, https://doi.org/10.1007/s10712-008-9038-y.

Hagemann, S., and T. Stacke, 2015: Impact of the soil hydrology scheme on simulated soil moisture memory. *Climate Dyn.*, **44**, 1731–1750, https://doi.org/10.1007/s00382-014-2221-6.

Humphrey, V., and L. Gudmundsson, 2019: GRACE-REC: A reconstruction of climate-driven water storage changes over the last century. *Earth Sys. Sci. Data*, **11**, 1153–1170, https://doi.org/10.5194/essd-11-1153-2019.

Jensen, L., A. Eicker, H. Dobslaw, T. Stacke, and V. Humphrey, 2019: Long-term wetting and drying trends in land water storage derived from GRACE and CMIP5 models. *J. Geophys. Res. Atmos.*, **124**, 9808–9823, https://doi.org/10.1029/2018JD029989.

Kim, H. J., 2017: Global soil wetness project phase 3 atmospheric boundary conditions (experiment 1). Data Integration and Analysis System (DIAS), accessed 22 September 2020, https://doi.org/10.20783/DIAS.501.

Kim, H.-M., P. J. Webster, and J. A. Curry, 2012: Evaluation of short-term climate change prediction in multi-model CMIP5 decadal hindcasts. *Geophys. Res. Lett.*, **39**, L10701, https://doi.org/10.1029/2012GL051644.

Kim, J., and B. D. Tapley, 2002: Error analysis of a low–low satellite-to-satellite tracking mission. *J. Guid. Control Dyn.*, **25**, 1100–1106, https://doi.org/10.2514/2.4989.

Kornfeld, R. P., B. W. Arnold, M. A. Gross, N. T. Dahya, W. M. Klipstein, P. F. Gath, and S. Bettadpur, 2019: GRACE-FO: The Gravity Recovery and Climate Experiment Follow-On Mission. *J. Spacecr. Rockets*, **56**, 931–951, https://doi.org/10.2514/1.A34326.

Liepert, B. G., and F. Lo, 2013: CMIP5 update of 'Inter-model variability and biases of the global water cycle in CMIP3 coupled climate models.' *Environ. Res. Lett.*, **8**, 029401, https://doi.org/10.1088/1748-9326/8/2/029401.

Luthcke, S. B., T. J. Sabaka, B. D. Loomis, A. A. Arendt, J. J. McCarthy, and J. Camp, 2013: Antarctica, Greenland and Gulf of Alaska land-ice evolution from an iterated GRACE global mascon solution. *J. Glaciol.*, **59**, 613–631, https://doi.org/10.3189/2013JoG12J147.

Marotzke, J., and Coauthors, 2016: MiKlip: A national research project on decadal climate prediction. *Bull. Amer. Meteor. Soc.*, **97**, 2379–2394, https://doi.org/10.1175/BAMS-D-15-00184.1.

Meehl, G. A., and Coauthors, 2009: Decadal prediction. *Bull. Amer. Meteor. Soc.*, **90**, 1467–1486, https://doi.org/10.1175/2009BAMS2778.1.

Mehrotra, R., A. Sharma, M. Bari, N. Tuteja, and G. Amirthanathan, 2014: An assessment of CMIP5 multi-model decadal hindcasts over Australia from a hydrological viewpoint. *J. Hydrol.*, **519**, 2932–2951, https://doi.org/10.1016/j.jhydrol.2014.07.053.

Müller, W. A., and Coauthors, 2012: Forecast skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology. *Geophys. Res. Lett.*, **39**, L22707, https://doi.org/10.1029/2012GL053326.

——, and Coauthors, 2018: A higher-resolution version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR). *J. Adv. Model. Earth Syst.*, **10**, 1383–1413, https://doi.org/10.1029/2017MS001217.

Pail, R., and Coauthors, 2015: Science and user needs for observing global mass transport to understand global change and to benefit society. *Surv. Geophys.*, **36**, 743–772, https://doi.org/10.1007/s10712-015-9348-9.

Palmer, T., R. Buizza, R. Hagedorn, A. Lawrence, M. Leutbecher, and L. Smith, 2006: Ensemble prediction: A pedagogical perspective. *ECMWF Newsletter*, No. 106, ECMWF, 10–17, http://doi.org/10.21957/AB129056EW.

Peel, M. C., B. L. Finlayson, and T. A. McMahon, 2007: Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci.*, **11**, 1633–1644, https://doi.org/10.5194/hess-11-1633-2007.

Pope, V. D., M. L. Gallani, P. R. Rowntree, and R. A. Stratton, 2000: The impact of new physical parametrizations in the Hadley Centre climate model: HadAM3. *Climate Dyn.*, **16**, 123–146, https://doi.org/10.1007/s003820050009.

Rodell, M., J. S. Famiglietti, D. N. Wiese, J. T. Reager, H. K. Beaudoing, F. W. Landerer, and M.-H. Lo, 2018: Emerging trends in global freshwater availability. *Nature*, **557**, 651–659, https://doi.org/10.1038/s41586-018-0123-1.

Scanlon, B. R., and Coauthors, 2018: Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. *Proc. Natl. Acad. Sci. USA*, **115**, E1080–E1089, https://doi.org/10.1073/pnas.1704665115.

Smith, D. M., and Coauthors, 2019: Robust skill of decadal climate predictions. *npj Climate Atmos. Sci.*, **2**, 13, https://doi.org/10.1038/s41612-019-0071-y.

Stacke, T., and S. Hagemann, 2016: Lifetime of soil moisture perturbations in a coupled land–atmosphere simulation. *Earth Syst. Dyn.*, **7** (1), 1–19, https://doi.org/10.5194/esd-7-1-2016.

Swart, N. C., and Coauthors, 2019: The Canadian Earth System Model version 5 (CanESM5.0.3). *Geosci. Model Dev.*, **12**, 4823–4873, https://doi.org/10.5194/gmd-12-4823-2019.

Swenson, S., J. Famiglietti, J. Basara, and J. Wahr, 2008: Estimating profile soil moisture and groundwater variations using GRACE and Oklahoma Mesonet soil moisture data. *Water Resour. Res.*, **44**, W01413, https://doi.org/10.1029/2007WR006057.

Syed, T. H., J. S. Famiglietti, M. Rodell, J. Chen, and C. R. Wilson, 2008: Analysis of terrestrial water storage changes from GRACE and GLDAS. *Water Resour. Res.*, **44**, W02433, https://doi.org/10.1029/2006WR005779.

Tapley, B. D., and Coauthors, 2019: Contributions of GRACE to understanding climate change. *Nat. Climate Change*, **9**, 358–369, https://doi.org/10.1038/s41558-019-0456-2.

Tatebe, H., and Coauthors, 2019: Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. *Geosci. Model Dev.*, **12**, 2727–2765, https://doi.org/10.5194/gmd-12-2727-2019.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1.

von Salzen, K., and Coauthors, 2013: The Canadian Fourth Generation Atmospheric Global Climate Model (CanAM4). Part I: Representation of physical processes. *Atmos.–Ocean*, **51**, 104–125, https://doi.org/10.1080/07055900.2012.755610.

Voss, K. A., J. S. Famiglietti, M. Lo, C. Linage, M. Rodell, and S. C. Swenson, 2013: Groundwater depletion in the Middle East from GRACE with implications for transboundary water management in the Tigris-Euphrates-western Iran region. *Water Resour. Res.*, **49**, 904–914, https://doi.org/10.1002/wrcr.20078.

Watanabe, M., and Coauthors, 2010: Improved climate simulation by MIROC5: Mean states, variability, and climate sensitivity. *J. Climate*, **23**, 6312–6335, https://doi.org/10.1175/2010JCLI3679.1.

Yuan, X., and E. Zhu, 2018: A first look at decadal hydrological predictability by land surface ensemble simulations. *Geophys. Res. Lett.*, **45**, 2362–2369, https://doi.org/10.1002/2018GL077211.

Zhang, L., H. Dobslaw, C. Dahle, I. Sasgen, and M. Thomas, 2016: Validation of MPI-ESM decadal hindcast experiments with terrestrial water storage variations as observed by the GRACE satellite mission. *Meteor. Z.*, **25**, 685–694, https://doi.org/10.1127/metz/2015/0596.

——, ——, T. Stacke, A. Güntner, R. Dill, and M. Thomas, 2017: Validation of terrestrial water storage variations as simulated by different global numerical models with GRACE satellite observations. *Hydrol. Earth Syst. Sci.*, **21**, 821–837, https://doi.org/10.5194/hess-21-821-2017.

Zhu, E., X. Yuan, and A. W. Wood, 2019: Benchmark decadal forecast skill for terrestrial water storage estimated by an elasticity framework. *Nat. Commun.*, **10**, 1237, https://doi.org/10.1038/S41467-019-09245-3.

Supplemental Material: Predictive skill assessment for land water storage in CMIP5 decadal hindcasts by a global reconstruction of GRACE satellite data

L. Jensen[1], **A. Eicker**[1], **T. Stacke**[2], **and H. Dobslaw**[3]

[1]Geodesy and Geoinformatics, HafenCity University, Hamburg, Germany
[2]Helmholtz-Zentrum Geesthacht, Centre for Materials and Coastal Research, Geesthacht, Germany
[3]Helmholtz Centre Potsdam, German Research Centre for Geosciences (GFZ), Potsdam, Germany

S1: Comparison of GRACE and GRACE-REC

Figure 1 shows the yearly global mean anomaly time series of GRACE-REC (Humphrey & Gudmundsson, 2019), as displayed in figure 1 of the main text, together with the original GRACE observations (Luthcke et al., 2013) used for creation of the GRACE-REC data set for the time span 1970 – 2014 (2003 – 2014 for observations). The GRACE observations lie within the error bounds of the reconstruction and the correlation of the two time series is 0.92. Thus, we consider GRACE-REC as a reasonably reliable proxy within the realm of our study.

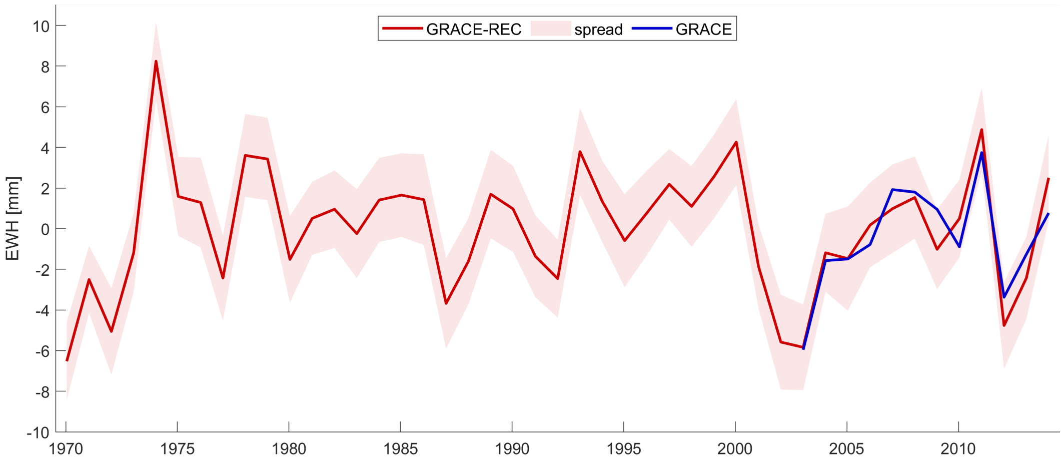


Figure 1. Yearly global mean anomaly time series of GRACE-REC (red) and original GRACE observations (blue) for the time span 1970 – 2014 (2003 – 2014 for observations).

Corresponding author: Laura Jensen, laura.jensen@hcu-hamburg.de

S2: Calculation of uncertainties

Along with the GRACE-REC time series $o(t)$, Humphrey and Gudmundsson (2019) provide an ensemble of randomly generated ensemble members $y_i(t)$ to incorporate the spatial and temporal error structure of the gridded TWS reconstruction into uncertainty estimates of spatial averages (e.g. global continental mean). The standard deviation of the GRACE-REC time series is computed as the unbiased sample standard deviation of these perturbed ensemble members $y_i(t)$ according to

$$\sigma_o(t) = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(y_i(t) - \bar{y}(t))^2}. \tag{1}$$

Here, M is the number of ensemble members (here 100), $y_i(t)$ is the TWS anomaly of the GRACE-REC member i at time step t, and $\bar{y}(t)$ is the arithmetic mean of all ensemble members.

The Init and Hist multi-model mean (MMM) time series (see figure 1 of the main text) are each calculated as the weighted mean of all ensemble members, to avoid that a single model contributes to the MMM with a higher weight due to a larger number of ensemble members:

$$m(t) = \frac{\sum_{i=1}^{N} w_i x_i(t)}{\sum_{i=1}^{N} w_i} = \frac{1}{V}\sum_{i=1}^{N} w_i x_i(t) \tag{2}$$

The total number of ensemble members is denoted by N (e.g. 39 for Init), $x_i(t)$ is the mTWS value of ensemble member i at time step t, and w_i is the weight assigned to the ensemble member i. The weights $w_i = 1/K$ are calculated as the reciprocal value of the number K of ensemble members per model, with K varying between 3 and 10. E.g., if a model has 3 members, each of them gets a weight of 1/3. As a result, all weights w_i sum up to the number of models $V = \sum_{i=1}^{N} w_i$ (e.g. 5).

When calculating the uncertainties of the MMM time series also the internal model uncertainties have to be taken into account, i.e., the weighted standard deviation has to be applied. Analogously to the weighted mean, the (biased) weighted standard deviation is given by

$$\sigma_x(t) = \sqrt{\frac{1}{V}\sum_{i=1}^{N} w_i (x_i(t) - m(t))^2}. \tag{3}$$

However, to obtain the *unbiased* weighted standard deviation the factor $\frac{1}{V}$ in equation 3 has to be adjusted. This is similar to the bias correction in the case of an unweighted standard deviation, where instead of $\frac{1}{M}$ the factor $\frac{1}{M-1}$ is applied (see equation 1). Bias correction for weighted standard deviation is not straight forward (Gatz & Smith, 1995), but it can be shown that

$$\sigma_x(t) = \sqrt{\frac{V}{V^2 - \sum_{i=1}^{N} {w_i}^2}\sum_{i=1}^{N} w_i (x_i(t) - m(t))^2} \tag{4}$$

is a good estimate for the *unbiased* weighted standard deviation, with the adjusted factor according to Kish (1965). Equation 4 describes the sample standard deviation for $x_i(t)$, i.e. the ensemble spread around the MMM, but not the standard deviation $\sigma_m(t)$ of the MMM $m(t)$ itself. Formally, $\sigma_m(t)$ can be derived via variance propagation of equation 2 utilizing the full variance-covariance matrix of the ensemble members $x_i(t)$. However, to come up with this covariance matrix, the error correlations between all members (of all models) have to be determined, which is not trivial. There is an ongoing discussion about the dependence of models and derived accuracies of multi-model averages in the climate modeling community (Knutti et al., 2010; Pennell & Reichler, 2011; Abramowitz & Bishop, 2015). In this study, we make the practical assumption of full error correlations between the members belonging to a particular model, and no error correlations between the members of different models. We admit that this is a simplification of the real error structure, however, the outlined uncertainty assessment can easily be adjusted if more realistic correlation estimates become available. With the current assumptions the standard deviation of the MMM becomes

$$\sigma_m(t) = \frac{1}{\sqrt{V}}\sigma_x(t). \tag{5}$$

The standard deviations of the correlations and RMSDs between the MMM and GRACE-REC provided as 1-sigma boundaries in figures 2, 3, 4, and 6 of the main text, are obtained via variance propagation from the standard deviations $\sigma_m(t)$ of the Init/Hist time series and $\sigma_o(t)$ of the GRACE-REC data set. Given that the bias of the time series is removed and thus the temporal mean is zero, the correlation of the MMM time series $m(t)$ and the GRACE-REC observational record $o(t)$ is

$$\rho = \frac{s_{mo}}{s_m \cdot s_o}, \tag{6}$$

where

$$s_{mo} = \textstyle\sum_{t=1}^{T} m(t)o(t), \quad s_m = \sqrt{\textstyle\sum_{t=1}^{T} m(t)^2}, \quad \text{and} \quad s_o = \sqrt{\textstyle\sum_{t=1}^{T} o(t)^2}, \tag{7}$$

and T denoting the length of the yearly time series (41 years). The standard deviation σ_ρ of the correlation is calculated by inserting equations 7 into equation 6 and performing variance propagation with the covariance matrix of $m(t)$ and $o(t)$ obtained from $\sigma_m(t)$ and $\sigma_o(t)$. As our time series has a low temporal resolution of yearly averages, we assume the error correlations between subsequent time steps to be negligible and set the corresponding elements in the covariance matrix to zero. If any reasons for assuming different correlations should arise, these can easily be adopted within this framework.

The same approach as described for the correlations is applied for the derivation of the standard deviations of the RMSD. The RMSD is calculated with

$$RMSD = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(m(t) - o(t))^2}. \tag{8}$$

and by propagating the uncertainties of $m(t)$ and $o(t)$ we obtain its standard deviation σ_{RMSD}.

S3: Identification of regions with incompatibilities between TWS from models and observations

We are aware of three geophysical signals contained in GRACE-derived TWS (and thus to some extent in the GRACE-REC data set) that are not explicitly represented in the ESMs:

- surface water variability (S)
- anthropogenic groundwater abstraction and irrigation (G)
- mass displacement due to large earthquakes (E)

We note that also natural groundwater variability is a part of GRACE TWS which is probably not properly represented in the climate models, but we would like to emphasize it is difficult (1) to separate the groundwater signal from the remaining compartments in the observations, and (2) to assess the degree to which groundwater is implicitly contained in the variable *total soil moisture content* of an ESM. We therefore focus in the following exclusively on the S/G/E effects. In regions where the influence of S/G/E is particularly large, TWS and mTWS are probably not entirely compatible which leads to a degradation of the results. We thus aim to identify such regions in order to exclude them from the further analysis.
In the following we briefly describe the data sets used and processing applied for the estimation of the magnitude of S/G/E effects and the approach for deriving the mask of regions to be excluded:

Surface water variability
Within the realm of the Research Unit GlobalCDA (funded by the German Research Foundation) a data set was produced which describes the monthly (2003/01 – 2016/12) mean influence of surface water storage change in lakes and reservoirs (in total 283 lakes/reservoirs obtained from the DAHITI data base) on the GRACE TWS signal (Deggim et al., manuscript in preparation). For this data set the surface water extent (from remote sensing) was combined with surface water level time series (from satellite altimetry) and converted to the spatial resolution of the GRACE data by applying appropriate spatial filtering (DDK3 filter, Kusche, 2007). We project the global 0.5° maps of this data set to 2° resolution and calculate annual anomalies. Afterward, we compute the root mean square (RMS) over 2003 – 2016 (figure 2a). We interpret this as the local influence of the annual surface water variability on the GRACE observations.

Anthropogenic groundwater abstraction
To estimate the magnitude of groundwater abstraction we make use of data from the hydrological model WaterGAP 2.2a (Döll et al., 2014). Net abstraction in WaterGAP 2.2a is defined as groundwater withdrawals minus return flow from irrigation with both surface water and groundwater. The global grids are converted from rates (in m^3/month) to monthly cumulated storage changes (EWH in mm), averaged per year, and remapped to 2° spatial resolution. Subsequently, the RMS over 1996 – 2009 (14 years) is calculated from the annual anomalies. To estimate the influence of these net abstraction changes on the GRACE observations we apply a GRACE-like spatial filtering (DDK3 filter, Kusche, 2007) to the resulting map of net abstraction RMS (figure 2b). Compared to the magnitude of surface water variability the RMS of the net abstraction is substantially smaller. This is due to the fact that anthropogenic groundwater abstraction mainly occurs as a linear mass trend whereas year-to-year variations are minor. Thus, although the regions generally affected by groundwater abstraction are of considerable size (Taylor et al., 2013), for the results of this study that focuses on annual anomalies excluding linear trends, groundwater abstraction only has a minor influence.

Mass displacement due to large earthquakes
Large earthquakes involve mass displacements detectable with GRACE that in first approximation cause a step function in the GRACE-derived time series of mass variations (with the discontinuity at the time of the earthquake). This step function cannot be removed by subtracting bias and linear trend as is done for the creation of the GRACE-REC data set. Thus, these mass variations distort the TWS estimates of GRACE-REC in earthquake regions. In-

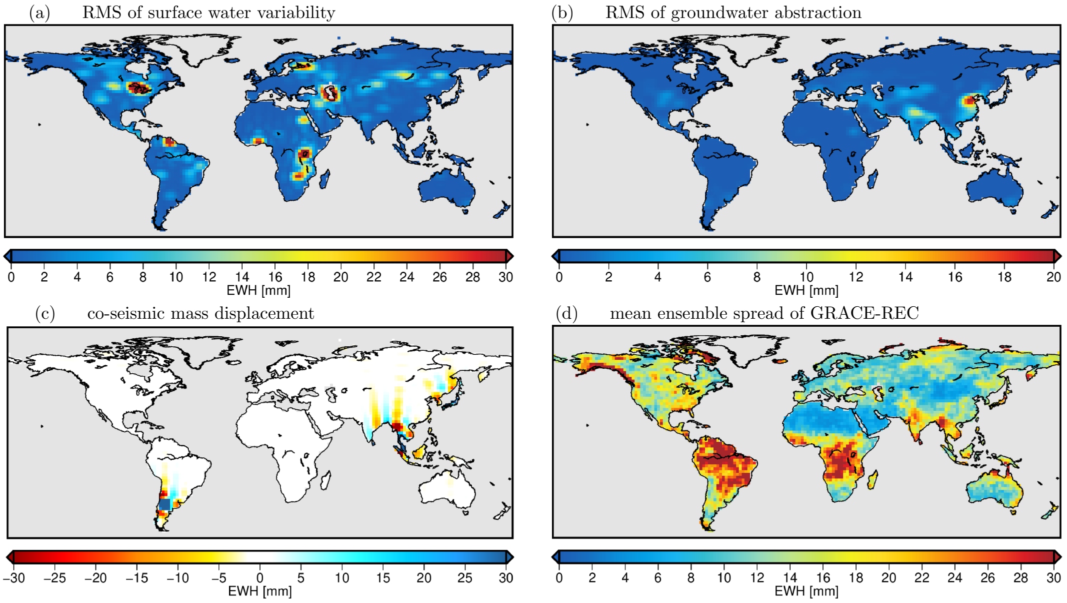


Figure 2. (a) RMS of annual surface water anomalies from 2003 – 2016 (b) RMS of annual net groundwater abstraction anomalies from 1996 – 2009 (c) co-seismic mass displacement due to Sumatra (2004), Chile (2010) and Tohoku (2011) earthquakes (d) mean standard deviation of 100 ensemble members of GRACE-REC annual anomalies between 2001 – 2014. The different time spans are due to data availability, but for comparability they all last 14 years (approx. GRACE time span).

formation on the spatial extent and magnitude of co-seismic mass variations contained in GRACE observations is, e.g., provided by Mayer-Gürr et al. (2018). They co-estimate the mass variations of the three largest earthquakes during the GRACE period (the Sumatra-Andaman 2004, the Chile 2010 and the Tohoku 2011 earthquake) together with the gravity field model ITSG-Grace2018s. The data are provided as spherical harmonic coefficients of the gravitational potential, and we convert them into EWH, apply a DDK3 filter (Kusche, 2007), and evaluate them on a 2° spatial grid (figure 2c). The assessment is restricted to the three largest earthquakes because GRACE is only sensitive to earthquakes with a magnitude of about > 8.5 (Pail et al., 2015; Han et al., 2013) and only these three earthquakes (with magnitudes 9.1, 8.8, and 9.0) are clearly above this threshold.

Derivation of a mask with regions to exclude

The magnitude of S/G/E effects (figure 2a-c) varies spatially and only substantially influences the GRACE-derived TWS in distinct regions. In order to identify these regions we use as a threshold for large S/G/E effects the noise floor of the GRACE-REC data set: For each grid cell the standard deviation of the 100 ensemble members of GRACE-REC is calculated for each year of the annual anomaly time series. We then average the standard deviations over the time span 2001 – 2014 (14 years) to obtain a mean observation spread (Figure 2d). The time span is slightly different to the GRACE time span (2003 – 2016) because the GRACE-REC data set ends in 2014. All grid cells where the RMS of the surface water variability, the groundwater abstraction, or the absolute influence of earthquakes (figure 2a-c) exceeds the spread of GRACE-REC (figure 2d) are excluded from the further analysis. These regions (figure 3) make up about 6.9% of the Earth's land surface (Greenland and Antarctica not considered).

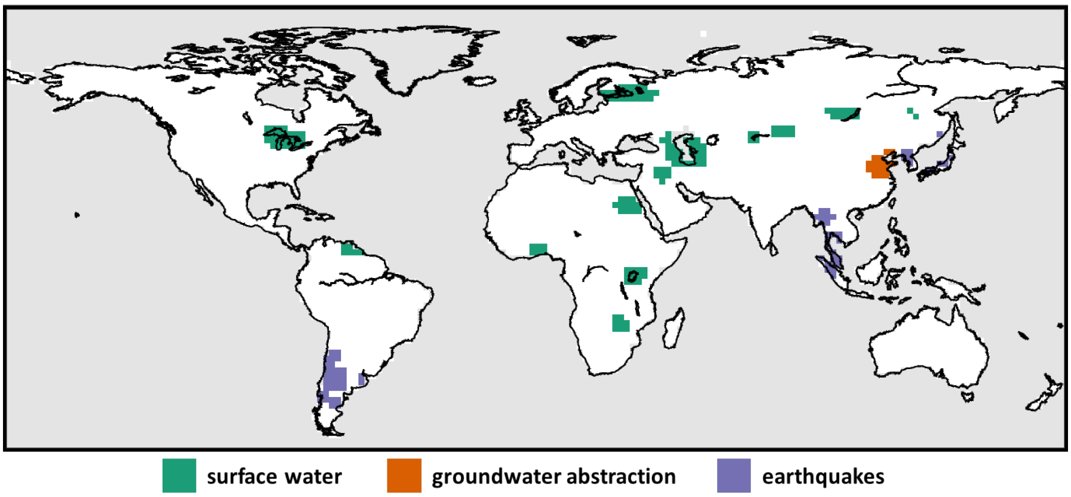


Figure 3. Mask with regions where the influence of surface water variability, anthropogenic groundwater abstraction, or the absolute mass displacement caused by earthquakes is larger than the spread of the GRACE-REC data set.

References

Abramowitz, G., & Bishop, C. H. (2015). Climate Model Dependence and the Ensemble Dependence Transformation of CMIP Projections. *Journal of Climate*, *28*(6), 2332–2348. (Publisher: American Meteorological Society) doi: 10.1175/JCLI-D-14-00364.1

Döll, P., Schmied, H. M., Schuh, C., Portmann, F. T., & Eicker, A. (2014). Global-scale assessment of groundwater depletion and related groundwater abstractions: Combining hydrological modeling with information from well observations and GRACE satellites. *Water Resources Research*, *50*(7), 5698–5720. doi: 10.1002/2014WR015595

Gatz, D. F., & Smith, L. (1995). The standard error of a weighted mean concentration—I. Bootstrapping vs other methods. *Atmospheric Environment*, *29*(11), 1185–1193. doi: 10.1016/1352-2310(94)00210-C

Han, S.-C., Riva, R., Sauber, J., & Okal, E. (2013). Source parameter inversion for recent great earthquakes from a decade-long observation of global gravity fields. *Journal of Geophysical Research: Solid Earth*, *118*(3), 1240–1267. (_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/jgrb.50116) doi: 10.1002/jgrb.50116

Humphrey, V., & Gudmundsson, L. (2019). GRACE-REC: a reconstruction of climate-driven water storage changes over the last century. *Earth System Science Data Discussions*, 1–41. doi: https://doi.org/10.5194/essd-2019-25

Kish, L. (1965). *Survey sampling.* New York: Chichester : Wiley.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., & Meehl, G. A. (2010). Challenges in Combining Projections from Multiple Climate Models. *Journal of Climate*, *23*(10), 2739–2758. (Publisher: American Meteorological Society) doi: 10.1175/2009JCLI3361.1

Kusche, J. (2007). Approximate decorrelation and non-isotropic smoothing of time-variable GRACE-type gravity field models. *Journal of Geodesy*, *81*(11), 733–749. doi: 10.1007/s00190-007-0143-3

Luthcke, S. B., Sabaka, T. J., Loomis, B. D., Arendt, A. A., McCarthy, J. J., & Camp, J. (2013). Antarctica, Greenland and Gulf of Alaska land-ice evolution from an iterated GRACE global mascon solution. *Journal of Glaciology*, *59*(216), 613–631. doi: 10.3189/2013JoG12J147

Mayer-Gürr, T., Behzadpur, S., Ellmer, M., Kvas, A., Klinger, B., Strasser, S., & Zehentner, N. (2018). *ITSG-Grace2018 - Monthly, Daily and Static Gravity Field Solutions from GRACE.* GFZ Data Services. doi: 10.5880/ICGEM.2018.003

Pail, R., Bingham, R., Braitenberg, C., Dobslaw, H., Eicker, A., Güntner, A., … IUGG Expert Panel (2015, November). Science and User Needs for Observing Global Mass Transport to Understand Global Change and to Benefit Society. *Surveys in Geophysics*, *36*(6), 743–772. doi: 10.1007/s10712-015-9348-9

Pennell, C., & Reichler, T. (2011). On the Effective Number of Climate Models. *Journal of Climate*, *24*(9), 2358–2367. (Publisher: American Meteorological Society) doi: 10.1175/2010JCLI3814.1

Taylor, R. G., Scanlon, B., Döll, P., Rodell, M., van Beek, R., Wada, Y., … Treidel, H. (2013). Ground water and climate change. *Nature Climate Change*, *3*(4), 322–329. doi: 10.1038/nclimate1744