Helmholtz-Zentrum
Geesthacht
Zentrum für Material- und Küstenforschung

Original

Jensen, L.; Eicker, A.; Stacke, T.; Dobslaw, H.:
Predictive Skill Assessment for Land Water Storage in CMIP5 Decadal Hindcasts by a Global Reconstruction of GRACE Satellite Data.
In: Journal of Climate. Vol. 33 (2020) 21, 9497 - 9509
First published online by AMS: 05.10.2020

https://dx.doi.org/10.1175/JCLI-D-20-0042.1

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

© 2020 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

Brought to you by HELMHOLTZ-ZENTRUM GEESTHACHT | Unauthenticated | Downloaded 03/19/21 11:39 AM UTC

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

Brought to you by HELMHOLTZ-ZENTRUM GEESTHACHT | Unauthenticated | Downloaded 03/19/21 11:39 AM UTC

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

Brought to you by HELMHOLTZ-ZENTRUM GEESTHACHT | Unauthenticated | Downloaded 03/19/21 11:39 AM UTC

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 $2° \times 2°$ 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

Brought to you by HELMHOLTZ-ZENTRUM GEESTHACHT | Unauthenticated | Downloaded 03/19/21 11:39 AM UTC

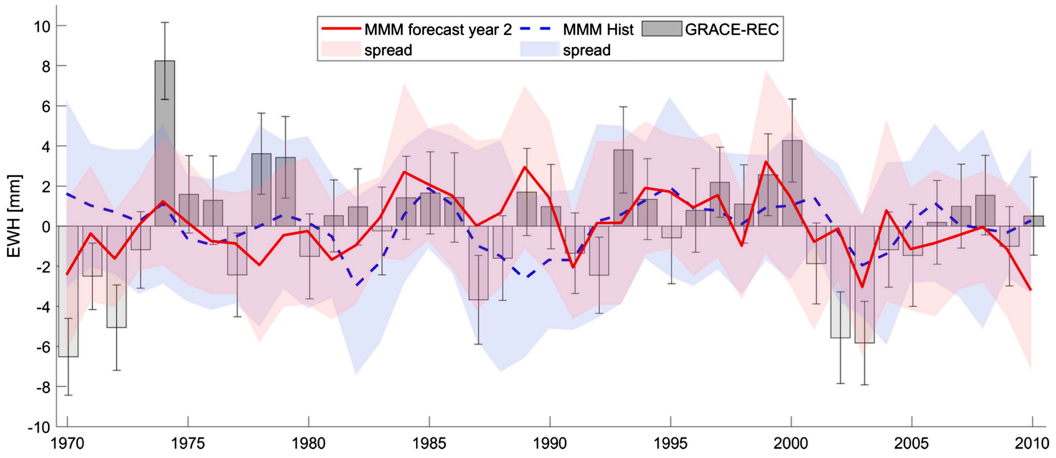

FIG. 1. Global land average (Greenland, Antarctica, and regions highly affected by surface waters, groundwater abstraction, and earthquakes excluded) annual time series for 1970–2010 for GRACE-REC TWS anomalies (gray bars), forecast year 2 initialized hindcast mTWS of multimodel mean (MMM; red line), 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

Brought to you by HELMHOLTZ-ZENTRUM GEESTHACHT | Unauthenticated | Downloaded 03/19/21 11:39 AM UTC

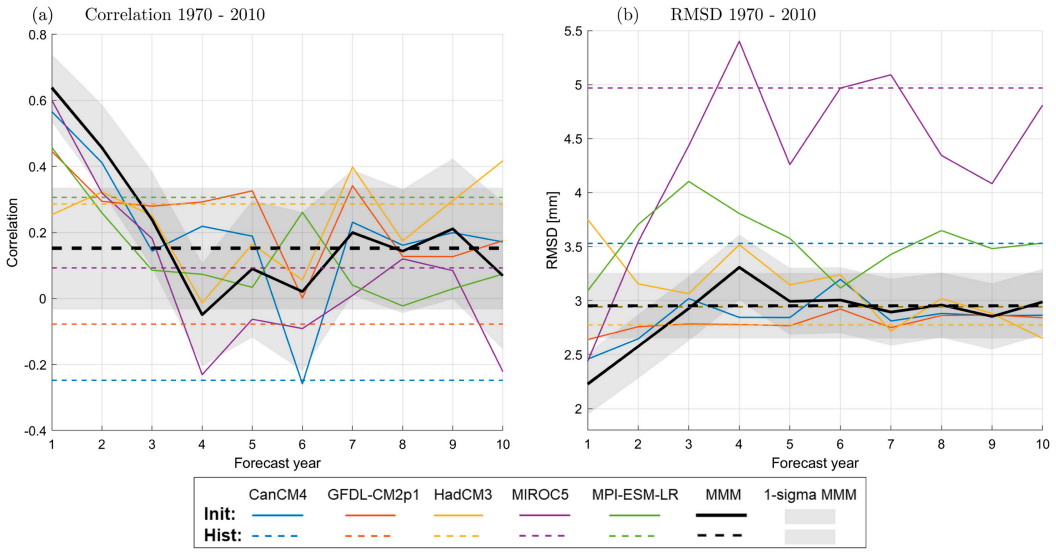

FIG. 2. (a) Correlations (as a function of forecast lead time) of the global mean GRACE-REC TWS anomaly time series with the global mean forecast year mTWS time series from decadal hindcasts (solid lines) and the uninitialized time series (dashed lines) for the time span 1970–2010. Colored lines indicate 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

Brought to you by HELMHOLTZ-ZENTRUM GEESTHACHT | Unauthenticated | Downloaded 03/19/21 11:39 AM UTC

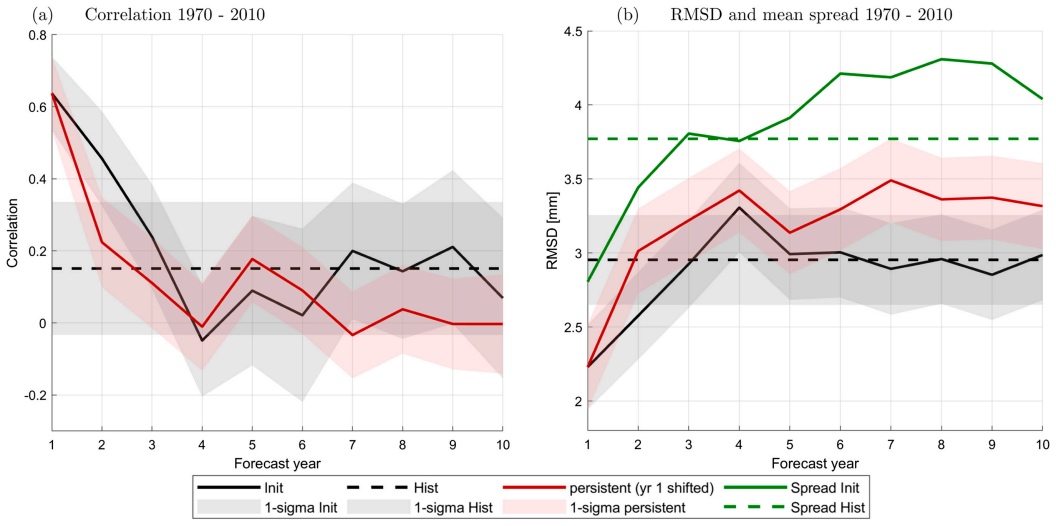

FIG. 3. Comparison of mTWS Init simulations (solid black, as in Fig. 2) and a persistent forecast (red) for the multimodel mean (MMM) global average time series for the time span 1970–2010 in terms of (a) correlation and (b) RMSD. The green lines in (b) denote the mean spread of the MMM for Init and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

Brought to you by HELMHOLTZ-ZENTRUM GEESTHACHT | Unauthenticated | Downloaded 03/19/21 11:39 AM UTC

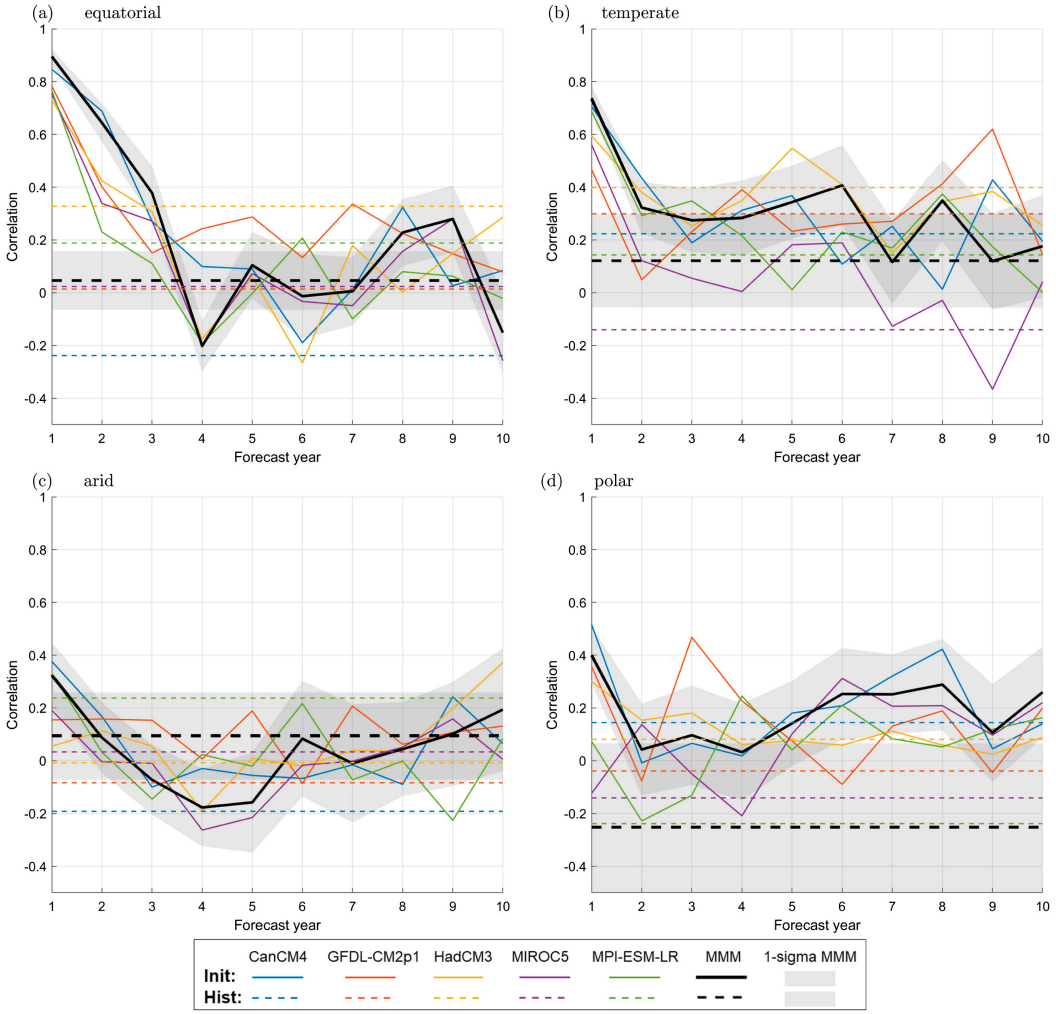

FIG. 4. Correlations (as a function of forecast lead time) of the GRACE-REC TWS anomaly time series with the forecast year mTWS time series from decadal hindcasts (solid lines) and the uninitialized time series (dashed lines) averaged over different climate zones for the time span 1970–2010. Colored 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.

Brought to you by HELMHOLTZ-ZENTRUM GEESTHACHT | Unauthenticated | Downloaded 03/19/21 11:39 AM UTC

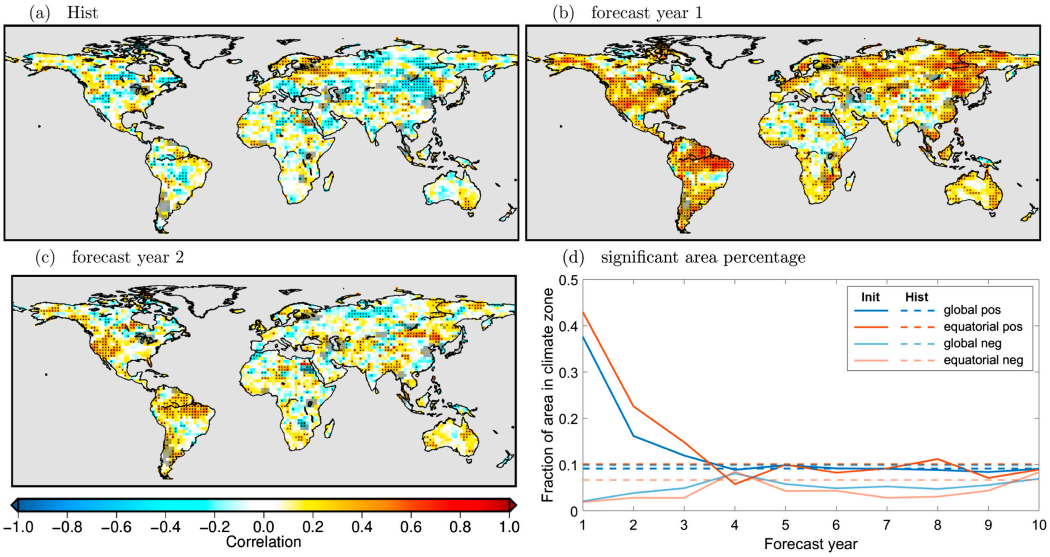

FIG. 5. Global maps of correlation of the GRACE-REC TWS anomaly time series with (a) uninitialized, (b) forecast year 1, and (c) forecast year 2 MMM mTWS anomaly time series for the time span 1970–2010. Stippled areas indicate significant correlation. Areas strongly affected by surface waters, groundwater 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

Brought to you by HELMHOLTZ-ZENTRUM GEESTHACHT | Unauthenticated | Downloaded 03/19/21 11:39 AM UTC

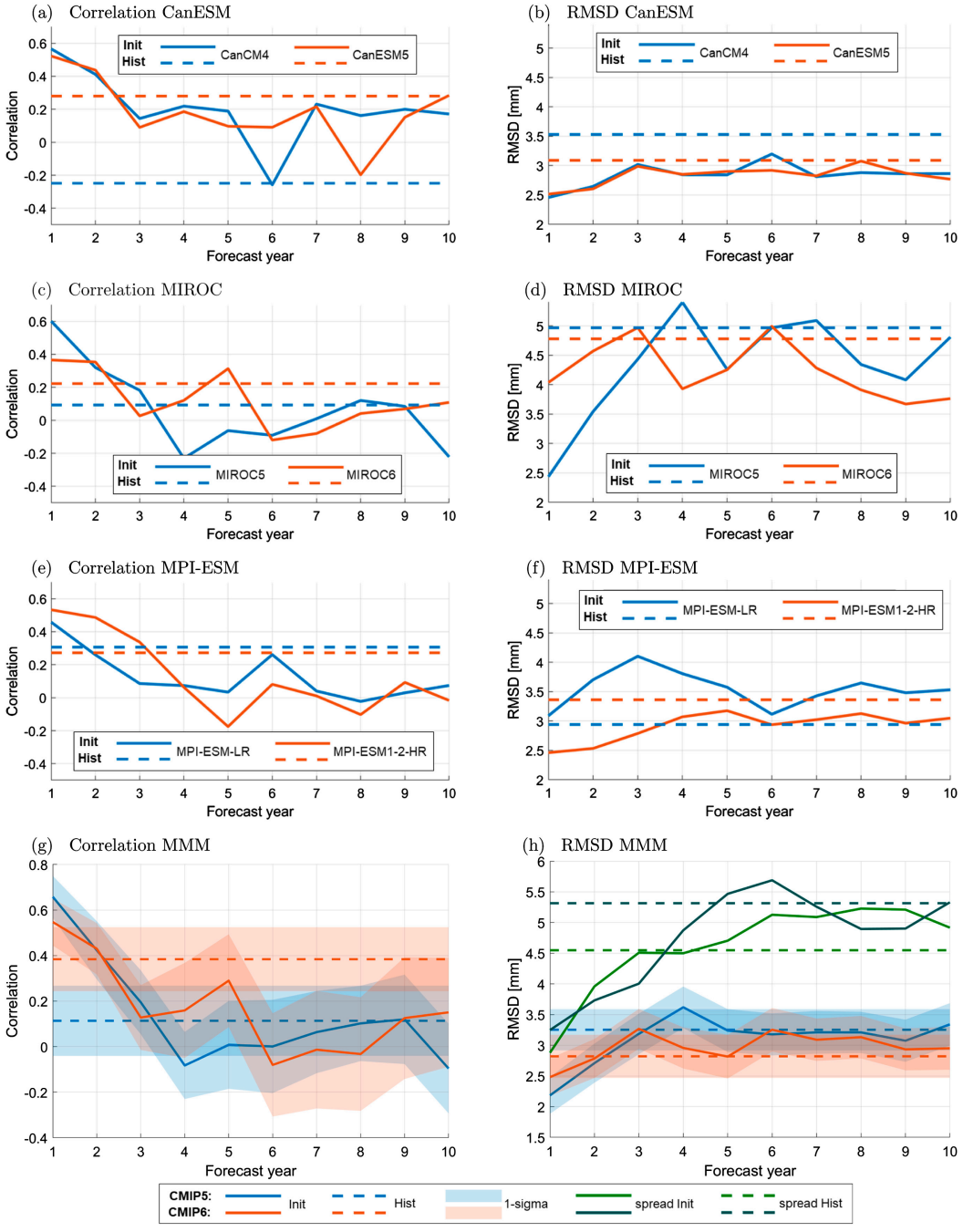

FIG. 6. Comparison of CMIP5 (blue) and CMIP6 (red) mTWS decadal hindcasts. (a),(c),(e) Correlation of the global mean GRACE-REC TWS time series and the Init and Hist mTWS time series as function of forecast time for three different ESMs for the time span 1970–2010. (g) The MMM from the three models above 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

Brought to you by HELMHOLTZ-ZENTRUM GEESTHACHT | Unauthenticated | Downloaded 03/19/21 11:39 AM UTC

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

Brought to you by HELMHOLTZ-ZENTRUM GEESTHACHT | Unauthenticated | Downloaded 03/19/21 11:39 AM UTC

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.

Brought to you by HELMHOLTZ-ZENTRUM GEESTHACHT | Unauthenticated | Downloaded 03/19/21 11:39 AM UTC

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.

Brought to you by HELMHOLTZ-ZENTRUM GEESTHACHT | Unauthenticated | Downloaded 03/19/21 11:39 AM UTC