Meteorologische Zeitschrift, Vol. **25**, No. 6, 685–694 (published online August 31, 2015)

MiKlip

Validation of MPI-ESM Decadal Hindcast Experiments with Terrestrial Water Storage Variations as Observed by the GRACE Satellite Mission

LIANGJING ZHANG[1*], HENRYK DOBSLAW[1], CHRISTOPH DAHLE[1], INGO SASGEN[1] and MAIK THOMAS[1,2]

[1]Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, 14473 Potsdam, Germany
[2]Free University of Berlin, Institute of Meteorology, Germany

(Manuscript received February 28, 2014; in revised form April 21, 2015; accepted April 25, 2015)

Abstract

Time-variations in the gravity field as observed by the GRACE mission provide for the first time quantitative estimates of the terrestrial water storage (TWS) at monthly resolution over one decade (2002–2011). TWS from GRACE is applied here to validate three different ensemble sets of decadal hindcasts performed with the coupled climate model MPI-ESM within the German research project MiKlip. Those experiments differ in terms of the applied low (LR) and medium (MR) spatial resolution configuration of MPI-ESM, as well as by the applied ensemble initialization strategy, where ocean-only (b0) is replaced by atmosphere and ocean (b1) anomaly initialization. Moderately positive skill scores of the initialized hindcasts are obtained both with respect to the zero anomaly forecast and the uninitialized projections in particular for lead year 1 in moderate to high latitudes of the Northern Hemisphere. Skill scores gradually increase when moving from b0-LR to b1-LR, and less prominent also for b1-LR to b1-MR, thereby documenting improvements of the MPI-ESM decadal climate prediction system during the most recent years.

Keywords: decadal climate hindcast validation, GRACE satellite data, terrestrial water storage, snow mass, soil moisture

1 Introduction

Reliable predictions of near-future changes in the Earth's climate beyond the seasonal time-scale would be highly valuable for the human society. Since adaption to and mitigation of climate change require far-reaching decisions many years before any measure might be effective, reliable estimates on low-frequency climate variability including near-term trends with lead time of a few years would be very beneficial for policy management and decision-making on climate change investments. This includes, for example, predicting the probability of extreme events as demonstrated in skillful multi-year predictions of Atlantic hurricane frequency (SMITH et al., 2010), but also near-surface air temperatures and the global water cycle.

The assessment of forecast skill of any decadal prediction system is typically based on validating extensive sets of hindcast experiments, where observations are already available (SMITH et al., 2007; KEENLYSIDE et al., 2008). Besides atmospheric and oceanic state quantities, also land surface conditions are increasingly recognized as influential for the evolution of climate in particular on the longer time scales. Water stored on the continents does not only affect the atmospheric circulation by means of surface albedo changes and thermal isolation due to snow cover, but also influences evaporation (KOSTER et al., 2004; MEEHL et al., 2009; SENEVIRATNE and STÖCKLI, 2007). Further, groundwater has a potentially large impact on the low-frequency climate variability by means of its contributions to soil moisture recharge (BIERKENS and VAN DEN HURK, 2007).

For a thorough skill assessment of any decadal prediction system it is therefore important not to focus solely on traditionally well-observed quantities as air temperatures or precipitation, but to take additionally many more types of observations into account (MAHMOOD et al., 2010). The satellite mission Gravity Recovery and Climate Experiment (GRACE; TAPLEY et al., 2004) provides since 2002 monthly snap-shots of the time-variable global gravity field of the Earth. Over the continents, gravity variations are primarily related to changes of water stored as snow, soil moisture, surface water, groundwater, and biomass. Since mass anomalies affect gravity measured by GRACE independently of a surface exposure, the GRACE experiment is the only realization of a remote sensing technique that is able to provide estimates of water masses integrated vertically from the surface down to the deep aquifers. The sum of the water masses in all those different hydrological storage compartments observed by GRACE is typically denoted as terrestrial water storage (TWS). TWS can be viewed as a measure of the vertical and lateral water fluxes in time and is therefore characterized by distinct low-frequency variability. This low-frequency character makes TWS ideally suit for decadal climate prediction

---

[*]Corresponding author: Liangjing Zhang, Helmholtz Centre Potsdam, GFZ German Research Centre for Geosciences, 14473 Potsdam, Germany, e-mail: liangjing.zhang@gfz-potsdam.de

DOI 10.1127/metz/2015/0596
Gebrüder Borntraeger Science Publishers, Stuttgart, www.borntraeger-cramer.com

model validation efforts. However, even though TWS estimates from GRACE are available for more than one decade up to now, the currently available knowledge on the overall accuracy of the GRACE-based TWS in particular on periods longer than seasonal is still limited, since it relies so far primarily on the assessment of hydrological model data (Syed et al., 2008) or the analysis of combined atmospheric-terrestrial water balances (Seneviratne et al., 2004).

In this study, we examine three different sets of decadal hindcasts performed with the coupled Earth System Model from Max-Planck-Institute for Meteorology (MPI-ESM) within the German research project "Mittelfristige Klimaprognosen" (MiKlip) (Müller et al., 2012; Pohlmann et al., 2013). We follow the verification framework proposed by Goddard et al. (2013) to assess the prediction quality of the MPI-ESM decadal hindcasts against terrestrial water storage anomalies from GRACE. Other satellite observations like the upper troposphere and lower stratosphere region (UTLS) temperature and satellite-retrieved cloud parameters have also been applied to evaluate the same MPI-ESM decadal hindcast experiments (Schmidt et al., accepted; Spangehl et al., 2015). The structure of our paper is arranged as follows: the processing of the GRACE satellite observations is outlined in Sec. 2. Subsequently, Sec. 3 gives a brief summary on the available hindcast experiments, whereas Sec. 4 sketches the validation approach applied. Results for deterministic skill scores are presented in Sec. 5, followed by some conclusions in the final section of this paper.

2 GRACE satellite observations

The U.S.-German twin satellite mission Gravity Recovery And Climate Experiment (GRACE) mission provides since April 2002 estimates of month-to-month changes in the gravitational field of the Earth mainly based on range-rate measurements between two low-flying satellites. After removing short-term variability due to tides in atmosphere (Biancale and Bode, 2006), solid earth (Petit and Luzum, 2010) and oceans (Savcenko and Bosch, 2012), as well as due to non-tidal variability in atmosphere and oceans (Dobslaw et al., 2013) from the observations, the resulting gravity changes represent mass transport phenomena in the Earth system, which are – apart from long-term trends – almost exclusively related to the global water cycle.

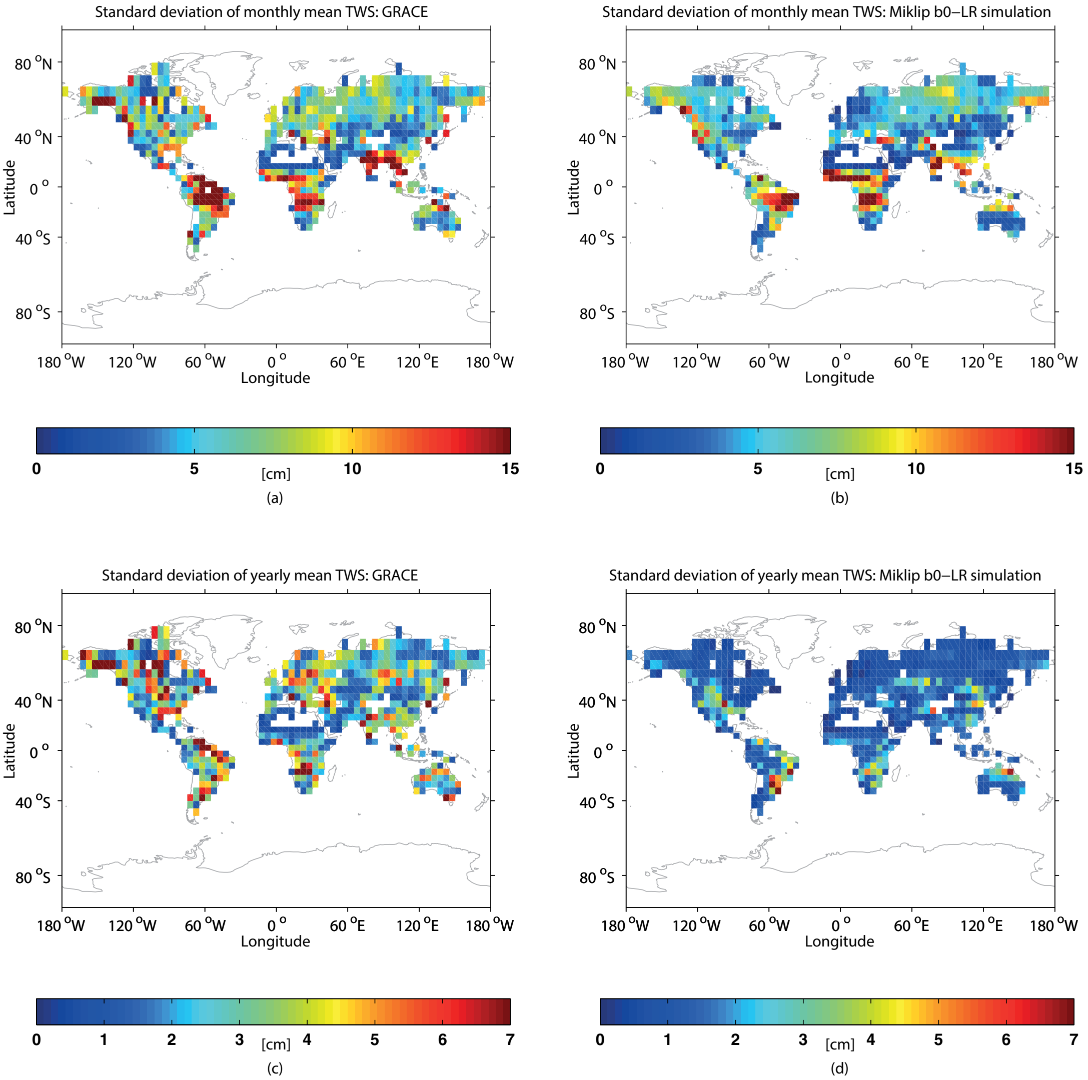
We use monthly GRACE RL05 Level-2 products from January 2003 to December 2011 processed at GFZ (Dahle et al., 2013), which are expressed in terms of fully normalized spherical harmonic coefficients up to degree and order 90, which approximately corresponds to a global resolution of 2 ° in latitude and longitude. We add global degree-1 coefficients as derived by Bergmann-Wolf et al. (2014), and remove a multi-year average covering the whole period to arrive at anomalies. Next, we apply the approximate decorrelation and non-isotropic smoothing method introduced by Kusche (2007) to remove spatially correlated errors in the gravity fields that are related to the particularly high sensitivity of GRACE along the direction of the satellite orbit. Smoothing in space domain is enabled through a tuning parameter of the signal covariance matrix that makes it approximately corresponding to a Gaussian filter radius of 240 km (Kusche et al., 2009). The filtered spherical harmonic coefficients that represent the anomalous gravity field of a particular month with respect to the multi-year average are synthesized into mass anomalies on a 1 ° regular grid following the conventions of Wahr et al. (1998).

Filtering may lead to biases, spatial leakage of signal and a further reduced spatial resolution. In order to account for this effect, local re-scaling factors are introduced following Klees et al. (2007). These scaling factors are estimated by applying all post-processing steps described above to TWS time-series from seven hydrological model experiments forced with re-analysis data, and regressing them onto the originally simulated TWS from the same model. The median value of the seven re-scaling factors per grid point is applied to the GRACE data, in order to arrive at robust estimates that are largely insensitive to deficiencies of any individual model. Finally, globally gridded and re-scaled TWS estimates from GRACE are aggregated into 5 ° averages as suggested by Goddard et al. (2013), and a mask is applied that removes ocean, small islands, and areas where the standard deviation of TWS as seen by GRACE is smaller than 1 cm of equivalent water height (eq.w.h.).

The time-variability in TWS as seen by GRACE is primarily dominated by the seasonal cycle (Figure 1a). We note standard deviation values of up to 15 cm eq.w.h. in tropical areas as, e.g., in the Amazon catchment, the Congo basin, and the Indian Monsoon area. Variabilitiy is substantially smaller at higher latitudes, but still reaches standard deviation values of up to 10 cm eq.w.h. in snow-dominated regions of North America and Eurasia. Calculating standard deviations from yearly averages (Figure 1c) in order to focus on differences between individual years only indicates much smaller variability of up to 5 cm eq.w.h. It is interesting to note that for the annual means the TWS variability at moderate latitudes is equally high as in tropical catchments.

3 Decadal hindcasts from MPI-ESM

Three different ensemble sets of decadal hindcasts are available from two different versions of the MPI-ESM coupled climate model. The low resolution (LR) model variant has been used already within the Coupled Model Intercomparison Project Phase 5 (CMIP5) (Taylor et al., 2012) and includes the oceanic component MPIOM (Jungclaus et al., 2013) discretized on a 1.5 ° curvilinear grid with 40 layers in the vertical, the atmospheric model ECHAM6 (Stevens et al., 2013) at T63 (1.875 °) horizontal resolution with 47 vertical levels that reach up to 0.1 hPa in the upper stratosphere,



(a) (b) (c) (d)

Figure 1: Standard deviation of monthly (top) and yearly averaged (bottom) TWS anomalies from GRACE (left) and the ensemble mean of a decadal hindcast experiment b0-LR (right) for lead year 1. Both data-sets are available from January 2003 to December 2011 and are aggregated here into 5 ° grid cells.

and the land surface model JSBACH (Reick et al., 2013; Hagemann et al., 2013). A second variant of MPI-ESM with medium resolution (MR) has a finer horizontal resolution in the ocean (0.4 ° curvilinear grid), and 95 levels in the atmosphere.

For both model variants, historical runs in line with the CMIP5 protocol are available until 2006. For the subsequent years, projections following the RCP4.5 emission scenario are used. In the remainder of this study, we denote those runs as ‘uninitialized projections’. In total, three ensemble members of projections are available from both versions of MPI-ESM. The anomaly technique (Pierce et al., 2004; Smith et al., 2013) is applied to initialize the baseline0 (b0) hindcasts with the oceanic temperature and salinity anomaly fields taken from a stand-alone simulation with the ocean model MPIOM forced with NCEP/NCAR atmospheric reanalysis data (Kalnay et al., 1996). Baseline1 (b1) instead uses atmospheric fields from ECMWF’s reanalysis ERA-Interim (Dee et al., 2011), and the oceanic anomaly fields from the Ocean Reanalysis System 4 (ORAS4) reanalysis (Balmaseda et al., 2013).

In order to cover the same time span, GRACE observations in the period 2003–2011 have been applied, and the following sets of decadal hindcasts are considered in the subsequent analysis:

(i) baseline0 at low resolution (b0-LR): three ensemble members initialized every year between 1994 and 2000, ten ensemble members initialized every year between 2001 and 2010;

(ii) baseline1 at low resolution (b1-LR): ten ensemble members initialized every year between 1994 and 2010; and

(iii) baseline1 at medium resolution (b1-MR): five ensemble members initialized every year between 1994 and 2010.

For all model runs, the standard output variables ‘total soil moisture content’ (mrso) and ‘surface snow

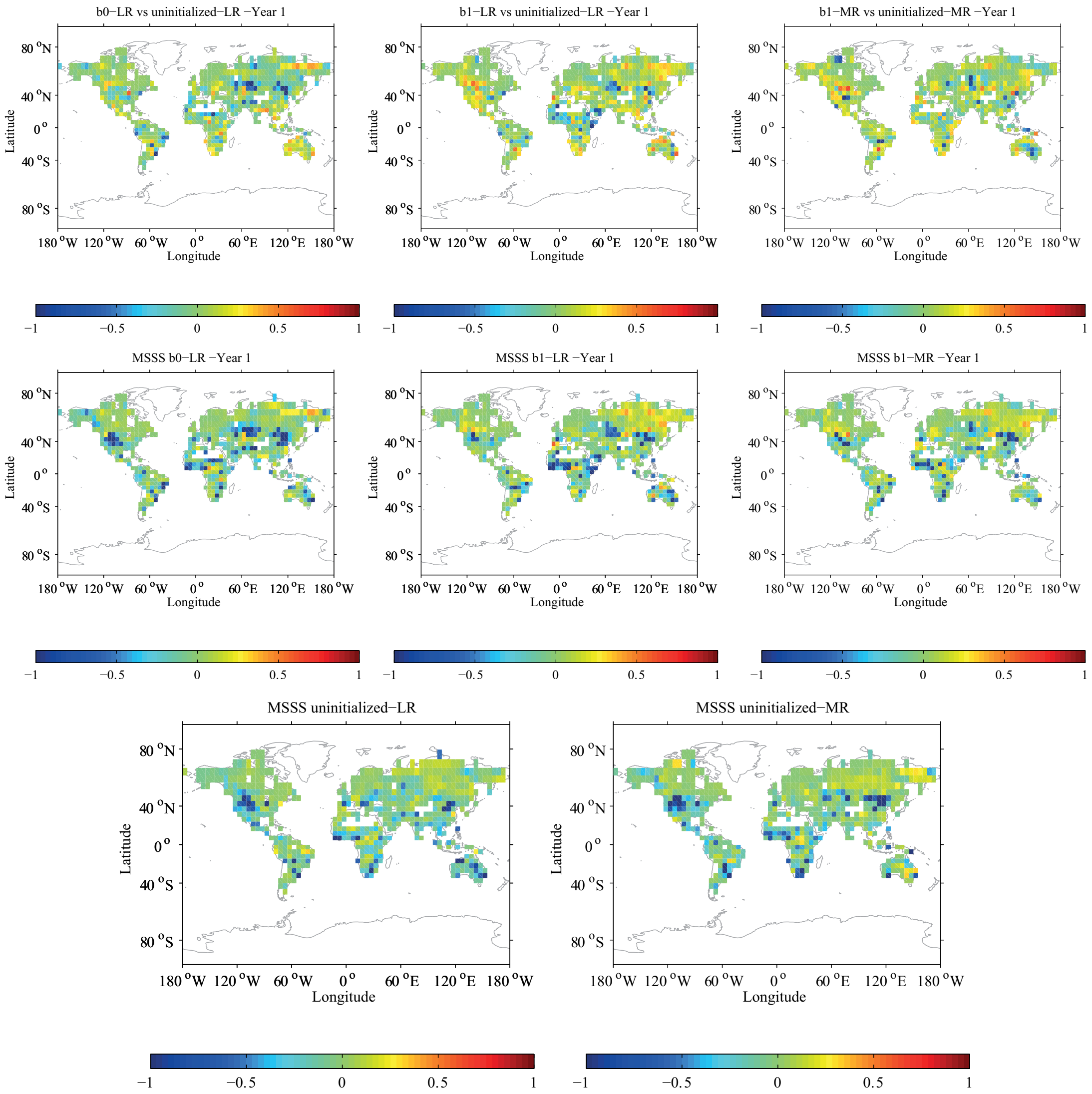


Figure 2: Mean squared skill score (MSSS) for three different decadal hindcast experiments within the MPI-ESM with respect to GRACE-based TWS observations for the period 2003–2011. Top row: MSSS of initialized hindcasts vs. uninitialized projections for lead year 1 of the experiments b0-LR (left), b1-LR (middle), and b1-MR (right). Middle row: MSSS of initialized hindcasts vs zero anomaly forecasts for lead year 1 of the experiments b0-LR (left), b1-LR (middle), and b1-MR (right). Bottom row: MSSS of uninitialized projections vs. zero anomaly forecasts of the uninitialized-LR (left), and the uninitialized-MR run.

amount' (snw) are taken as the dominant contributors to the terrestrially stored water; the sum of both is contrasted against GRACE-based TWS in the remainder of this study. Note that both surface water and deep aquifers are likely to contribute to TWS as well, their effects have been, however, omitted in this assessment, since none of them is properly represented by MPI-ESM. We are primarily interested in assessing the hindcast quality in relation to the lead time, thus we re-sort the data and build lead year time-series for the ensemble means of each ensemble set of hindcasts.

Exemplarily, variability of TWS as simulated within the b0-LR hindcast experiment is depicted in Figure 1b. In general, signal magnitudes and spatial patterns fit well to the GRACE observations. Boundaries between different hydro-climates are more sharply defined in the model results, suggesting that spatial leakage is still present in the post-processed and re-scaled GRACE results. For the standard deviation of annual means, however, we note substantially lower variability compared to the GRACE data in almost all regions of the world, and the magnitudes typically reach only half the amount of the observations (Figure 1d).

4 Validation approach

By using TWS from GRACE as the reference observations, we are going to assess how skillful the three different sets of hindcasts from the MPI-ESM are in terms of predicting anomalous terrestrial water storage. By posing this validation question, we attempt to investigate whether the models provide information about water availability below the long-term mean (i.e., which might lead to drought conditions) or above it (i.e., which might lead to flood conditions). Since the trends in TWS observed by GRACE cannot be exclusively assigned to natural climate variability, but are also due to, for instance, solid earth geophysics and anthropogenic groundwater extraction from deep aquifers, the average trends of TWS over the observation period are removed from both the observations and hindcasts. To exclude also seasonal effects, which are not in the focus of a decadal climate prediction system, we confine ourselves to the assessment on annual averages only. Finally, both model and observational data-sets are aggregated onto a regular 5 ° grid as used also in Goddard et al. (2013). As a deterministic metric, we use the mean-squared skill score (MSSS)

$$MSSS(P,R,O) = 1 - \frac{MSE(P,O)}{MSE(R,O)}, \tag{4.1}$$

that compares the mean-squared errors (MSE)

$$MSE(P,O) = \frac{1}{n}\sum_{j=1}^{n}(P_j - O_j)^2 \tag{4.2}$$

of two sets of predictions P and R with respect to the observations O (Murphy, 1988). This score is also implemented in the central evaluation system of MiKlip as described by Kadow et al. (2015). Here, P_j is the prediction for a certain time-step j, and O_j the corresponding observation made at the same time. Both P_j and O_j are anomalies relative to their respective climatologies calculated over the same data span. Besides the uninitialized projections, we also use 'zero-anomaly forecasts' as a reference R in Eq. (4.1). As a deterministic score, the MSE is typically calculated from ensemble mean averages that we obtain from the re-sorted lead year dependent time-series from the different sets of ensemble hindcasts as described in the previous section. Besides showing MSSS for the single lead year 1, average MSSS over lead years 2–5, and 6–9 are also calculated. Following the non-parametric bootstrapping method described in Appendix 2 of Goddard et al. (2013), we also assess the significance of the MSSS of initialized hindcasts b1-MR with respect to uninitialized projections for lead year 1 with three different significance levels of 90 %, 95 % and 99 %.

5 Results

With globally gridded TWS anomalies at 5 ° spatial resolution from GRACE as the reference, MSSS maps are calculated for the three different MPI-ESM decadal prediction ensembles. Focusing on lead year 1 (Figure 2), we note that the scores of the uninitialized projections compared to a zero-anomaly forecast are generally small for the LR and MR model version considered: estimates hardly exceed 0.5 even in isolated regions. Nevertheless, we note a slight improvement of the scores for MR in moderate to high latitudes of the Northern Hemisphere, and also a reduction of the highly negative skills shown in LR in the Southeast-Asian Monsoon regions. When looking into scores of the initialized runs with respect to a zero-anomaly forecast, we find moderate skills in large parts of Siberia in b0-LR that are substantially improved in b1-LR, indicating that the combined atmosphere/ocean initialization strategy implemented in baseline1 is also beneficial to the representation of land-surface processes in MPI-ESM. This improvement is also reflected in scores calculated for the initialized versus the uninitialized projections, that also increase from b0-LR to b1-LR, culminating in the latter in positive skill scores in almost all regions apart from the large deserts, where TWS variations are small and observations are certainly dominated by errors. Skills are also largely positive for b1-MR when comparing predictions to the uninitialized projections. However, since the uninitialized projections already show better skill than the LR projection with respect to a zero-anomaly forecast, there is no obvious improvement compared with b1-LR. We note from the significance figure (Figure 3) that the regions where we discussed changes in the skill scores (e.g., parts of Northern Eurasia, Monsoon areas) indeed show significant skills.

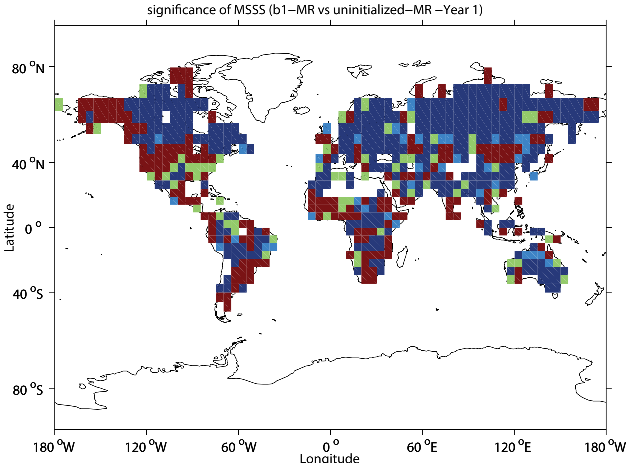


Figure 3: Significance of the MSSS of initialized hindcasts b1-MR with respect to uninitialized projections for lead year 1. Red areas shows statistical significance that the MSSS is positive at the 99 % confidence level, green indicates the added areas where the significance level is chosen as 95 %, light blue is for 90 % and dark blue represents the other areas.

For longer lead time of the predictions, however, skill scores of all three model runs diminish since the influence of the initialization becomes increasingly smaller. Skill scores averaged over lead years 2–5 are generally smaller when calculated against the zero-anomaly forecast (Figure 4), and the distinctness of the two different initialization strategies or model versions are not substantial anymore. With respect to the uninitialized projections, we still note generally positive skills for all three hindcast ensembles considered, but notable differences between baseline0 and baseline1, or LR and MR

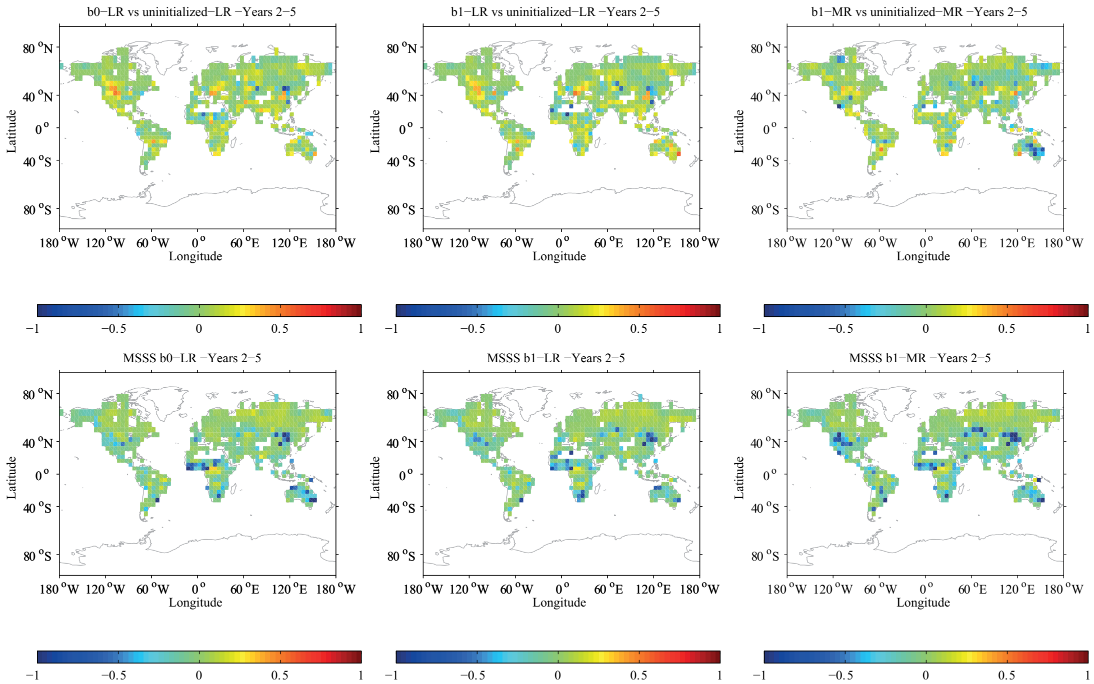


Figure 4: Same as Figure 2, but for lead years 2–5.

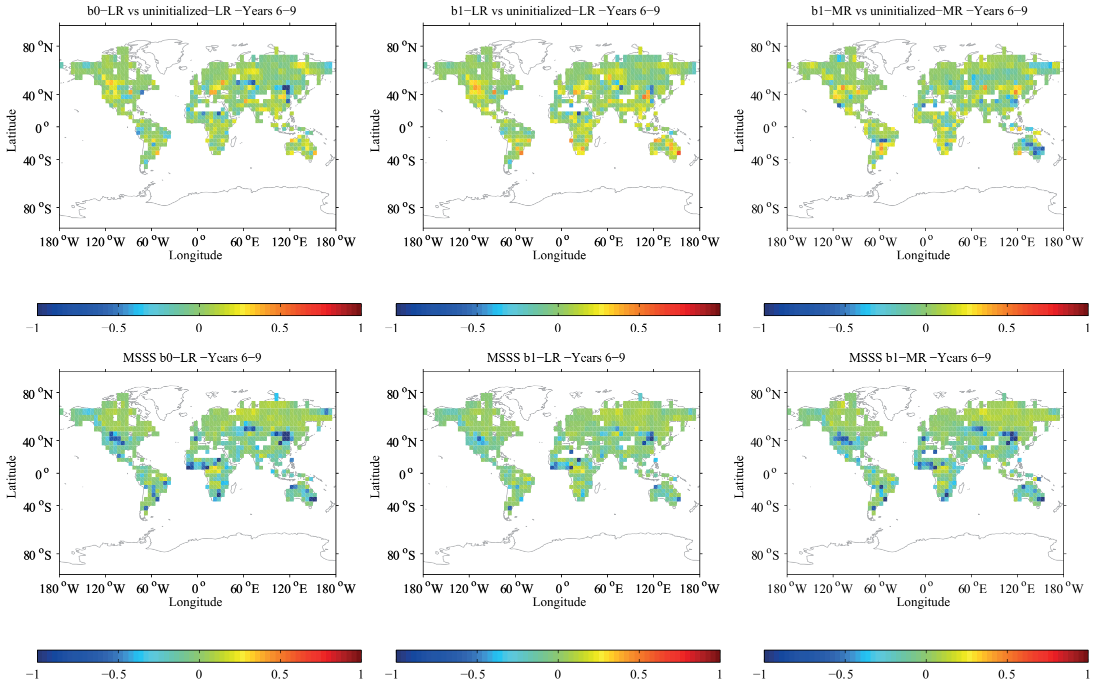


Figure 5: Same as Figure 2, but for lead years 6–9.

are no longer evident, and scores are regionally much less coherent than for lead year 1. Generally the same holds true for the averaged scores over lead years 6–9 (Figure 5). Although still positive scores are obtained for the initialized forecasts when compared to uninitialized projections, those are typically only slightly larger than zero and do show very little regional coherence.

To analyze the time-variability of the simulated terrestrial water storage for selected regions in more detail, we average TWS from both GRACE and the different model runs for two differently located discharge basins: The Lena catchment in Central Siberia whose climate is generally classified as 'snow dominated, fully humid with cool summer (Dfc)' according to the Köppen-Geiger Climate Classification (Kottek et al., 2006), and the Nelson River catchment in North America with a slightly different hydro-climate of 'snow dominated, fully humid with warm summer (Dfb)', in at least the

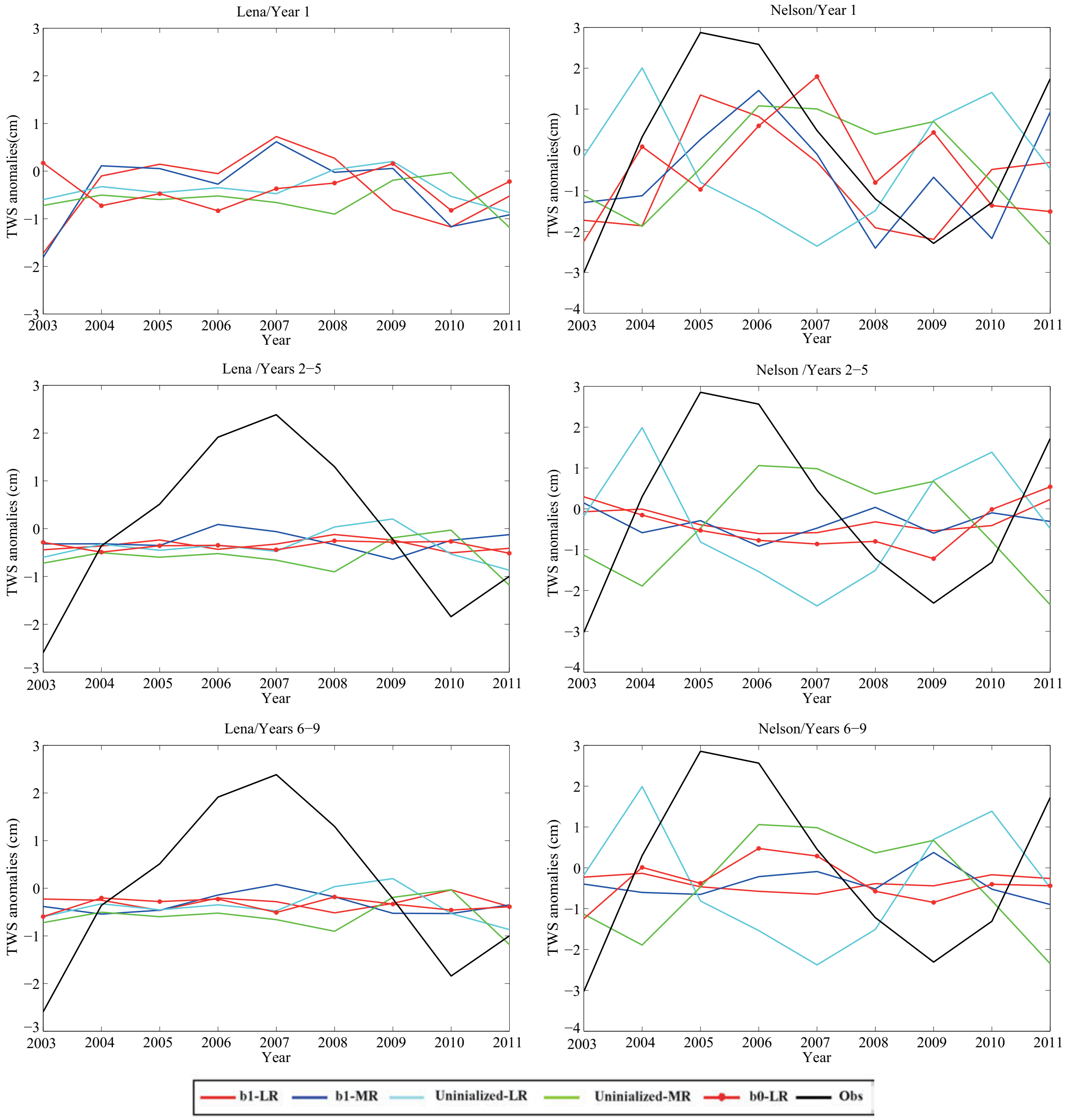


Figure 6: Time series of yearly averages of TWS for the Lena catchment in Siberia (left) and the Nelson catchment in North America (right): GRACE-based observations (black), uninitialized projection at low resolution (cyan), uninitialized projection at medium resolution (green), and decadal hindcast experiments b0-LR (red∗), b1-LR (red), and b1-MR (blue).

major fraction of its area. For the Lena catchment (Figure 6), we note for the lead year 1 a generally good representation of the TWS anomalies in the dry years 2003 and 2010, that are represented well in both b1-LR and b1-MR. The baseline0 hindcast, however, does not capture those anomalies, its time-variability instead is rather comparable with the uninitialized projections of both model versions. The wet anomaly in 2007 is also captured by both b1-LR and b1-MR, but with much smaller peak amplitudes. In particular, multi-year changes in TWS potentially related to storage changes in deeper soil layers and near-surface aquifers are not reflected in the MPI-ESM model runs. Since feedbacks from those deeper storage reservoirs to the climate system are generally difficult to quantify, it remains open at this point if this deficit of MPI-ESM is important for the quality of decadal climate predictions. Similar conclusions for lead year 1 are obtained for the Nelson catchment. We note a general increase in area-averaged TWS during the years 2003 to 2005 in both b1-LR and b1-MR, which closely resembles the observed variability but is not present in the uninitialized projections. We see a subsequent decline in TWS during the following years until 2008, which is again quite nicely captured by both baseline1 hindcast ensembles. Dry conditions in 2009, however, are only reproduced by b1-LR, whereas the MR hindcasts rather predict a rapid increase in TWS for that year. Focusing on longer lead times of 2–5 or 6–9 years, however, we note substantially smaller differences between hindcasts and projections for both model versions in both catchments considered. For those lead years, the ensemble mean of the hindcasts is more frequently close to a zero-anomaly forecast, supporting our conclusions drawn from Figures 4 and 5.

6 Discussion and conclusion

One decade of terrestrial water storage estimates from GRACE are applied for the validation of three different decadal hindcast experiments with the MPI-ESM coupled climate model. Mean squared skill scores of annual averages of the ensemble means indicate positive skill for the initialized hindcasts in particular for lead year 1 in moderate to high latitudes of the Northern Hemisphere when compared to the uninitialized simulations or their climatologies. In addition, skills gradually increase when moving from b0-LR to b1-LR, and also less pronounced from b1-LR to b1-MR, indicating that changes in the initialization and increased resolution implemented in the different experiments indeed lead to more skillful initialized hindcasts than in the earlier experiments.

For lead years greater than 1, however, skill rapidly drops down towards zero, and also the differences between the three experiments are diminished. It appears that the simulated TWS variability is substantially lower for the uninitialized projections (and the later lead years of the initialized runs), indicating that the models tend to stay closer to its climatology instead of simulating substantial deviations from it, which is, however, not uncommon in numerical modeling experiments of the climate system.

The average level of skill in predicting TWS in all three experiments is quite modest, in particular when compared to more traditionally considered validation variables. This might be related to the relatively short period of observations, but since GRACE is still in operation and a follow-on mission is scheduled to launch in 2017 (Flechtner et al., 2013), this situation will gradually improve during the next years. Further, contamination of GRACE results due to spatial leakage and mass transport processes not related to the global water cycle must be taken into account. In this regard, ongoing re-processing efforts and refined post-processing approaches will contribute to a more accurate quantification of random noise and systematic biases in the satellite data.

The good capture of the dry and wet anomalies in the first year of the initialized runs in some areas and the improvement caused by the new initialization strategy is encouraging, given that only oceanic and atmospheric initialization is applied. Earlier studies already demonstrated that land surface initialization is able to additionally contribute – via direct feedbacks into the atmosphere – to the skill of sub-seasonal and seasonal predictions by, for instance, increasing the boreal summer predictability from soil moisture observations (Dirmeyer, 2005; Koster et al., 2010; Douville, 2010). Since also multi-year memory effects of soil moisture have been found to have promising feedbacks on the climate (Seneviratne et al., 2013), it would be straightforward to also consider GRACE-based TWS as an observable to be assimilated into a land surface re-analysis data-set, that might be subsequently used for initialization of future decadal climate prediction experiments.

Acknowledgments

This study has been supported by the German Federal Ministry of Education and Research (BMBF) within the FONA research program under grants 03F0654A and 01LP1151A. We would also like to thank the German Space Operations Center (GSOC) of the German Aerospace Center (DLR) for providing continuously and nearly 100 % of the raw telemetry data of the twin GRACE satellites.

References

Balmaseda, M.A., K. Mogensen, A.T. Weaver, 2013: Evaluation of the ECMWF ocean reanalysis system ORAS4. – Quart. J. Roy. Meteor. **139**, 1132–1161, DOI: 10.1002/qj.2063.

Bergmann-Wolf, I., L. Zhang, H. Dobslaw, 2014: Global eustatic sea-level variations for the approximation of geocenter motion from GRACE. – J. Geod. Sci. **4**, DOI: 10.2478/jogs-2014-0006.

Biancale, R., A. Bode, 2006: Mean annual and seasonal atmospheric tide models based on 3-hourly and 6-hourly ECMWF surface pressure data. – Scientific Technical Report STR06/01, GFZ, Helmholtz-Zentrum, Potsdam.

Bierkens, M.F.P., B.J.J.M. van den Hurk, 2007: Groundwater convergence as a possible mechanism for multi-year persistence in rainfall. – Geophys. Res. Lett. **34**, published online, DOI: 10.1029/2006GL028396.

Dahle, C., F. Flechtner, C. Gruber, D. König, R. König, G. Michalak, K. Neumayer, 2013: GFZ GRACE Level-2 Processing Standards Document for Level-2 Product Release 0005. – Scientific Technical Report STR12/02 – Data, Revised Edition, GFZ, Helmholtz-Zentrum, Potsdam.

Dee, D.P., S.M. Uppala, A.J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M.A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A.C.M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A.J. Geer, L. Haimberger, S.B. Healy, H. Hersbach, E.V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A.P. McNally, B.M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, F. Vitart, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. – Quart. J. Roy. Meteor. **137**, 553–597, DOI: 10.1002/qj.828.

Dirmeyer, P.A., 2005: The land surface contribution to the potential predictability of boreal summer season climate. – J. Hydrometeor. **6**, 618–632, DOI: 10.1175/JHM444.1.

Dobslaw, H., F. Flechtner, I. Bergmann-Wolf, C. Dahle, R. Dill, S. Esselborn, I. Sasgen, M. Thomas, 2013: Simulating high-frequency atmosphere-ocean mass variability for dealiasing of satellite gravity observations: AOD1B RL05. – J. Geophys. Res. **118**, 3704–3711, DOI: 10.1002/jgrc.20271.

Douville, H., 2010: Relative contribution of soil moisture and snow mass to seasonal climate predictability: a pilot study. – Climate Dyn. **34**, 797–818, DOI: 10.1007/s00382-008-0508-1.

Flechtner, F., M. Watkins, P. Morton, F. Webb, 2013: Status of the GRACE Follow-on Mission. – In: Proceedings of the International Association of Geodesy Symposia Gravity, Geoid and Height System (GGHS2012), volume IAGS-D-12-00141, Venice, Italy.

Goddard, L., A. Kumar, A. Solomon, D. Smith, G. Boer, P. Gonzalez, V. Kharin, W. Merryfield, C. Deser, S. Mason, B. Kirtman, R. Msadek, R. Sutton, E. Hawkins, T. Fricker, G. Hegerl, C. Ferro, D. Stephenson, G. Meehl, T. Stockdale, R. Burgman, A. Greene, Y. Kushnir, M. Newman, J. Carton, I. Fukumori, T. Delworth, 2013: A verification framework for interannual-to-decadal predictions experiments. – Climate Dyn. **40**, 245–272, DOI: 10.1007/s00382-012-1481-2.

Hagemann, S., A. Loew, A. Andersson, 2013: Combined evaluation of MPI-ESM land surface water and energy fluxes. – J. Adv. Model. Earth Syst. **5**, 259–286, DOI: 10.1029/2012MS000173.

Jungclaus, J.H., N. Fischer, H. Haak, K. Lohmann, J. Marotzke, D. Matei, U. Mikolajewicz, D. Notz, J.S. von Storch, 2013: Characteristics of the ocean simulations in the Max Planck Institute Ocean Model (MPIOM) the ocean component of the MPI-Earth system model. – JAMES **5**, 422–446, DOI: 10.1002/jame.20023.

Kadow, C., S. Illing, O. Kunst, H. Rust, H. Pohlmann, W.A. Müller, U. Cubasch, 2015: Evaluation of forecasts by accuracy and spread in the MiKlip decadal climate prediction system. – Meteorol. Z. **25**, 631–644, DOI: 10.1127/metz/2015/0639.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, 1996: The NCEP/NCAR 40-year reanalysis project. – Bull. Amer. Meteor. Soc. **77**, 437–471, DOI: 10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

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

Klees, R., E.A. Zapreeva, H.C. Winsemius, H.H.G. Savenije, 2007: The bias in GRACE estimates of continental water storage variations. – Hydrol. Earth Syst. Sc. **11**, 1227–1241, DOI: 10.5194/hess-11-1227-2007.

Koster, R.D., P.A. Dirmeyer, Z. Guo, G. Bonan, E. Chan, P. Cox, C.T. Gordon, S. Kanae, E. Kowalczyk, D. Lawrence, P. Liu, C.-H. Lu, S. Malyshev, B. McAvaney, K. Mitchell, D. Mocko, T. Oki, K. Oleson, A. Pitman, Y.C. Sud, C.M. Taylor, D. Verseghy, R. Vasic, Y. Xue, T. Yamada, 2004: Regions of strong coupling between soil moisture and precipitation. – Science **305**, 1138–1140, DOI: 10.1126/science.1100217.

Koster, R.D., S.P.P. Mahanama, T.J. Yamada, G. Balsamo, A.A. Berg, M. Boisserie, P.A. Dirmeyer, F.J. Doblas-Reyes, G. Drewitt, C.T. Gordon, Z. Guo, J.-H. Jeong, D.M. Lawrence, W.-S. Lee, Z. Li, L. Luo, S. Malyshev, W.J. Merryfield, S.I. Seneviratne, T. Stanelle, B.J.J.M. van den Hurk, F. Vitart, E.F. Wood, 2010: Contribution of land surface initialization to subseasonal forecast skill: First results from a multi-model experiment. – Geophys. Res. Lett. **37**, DOI: 10.1029/2009GL041677.

Kottek, M., J. Grieser, C. Beck, B. Rudolf, F. Rubel, 2006: World Map of the Köppen-Geiger climate classification updated. – Meteorol. Z. **15**, 259–263, DOI: 10.1127/0941-2948/2006/0130.

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

Kusche, J., R. Schmidt, S. Petrovic, R. Rietbroek, 2009: Decorrelated GRACE time-variable gravity solutions by GFZ, and their validation using a hydrological model. – J. Geodesy **83**, 903–913, DOI: 10.1007/s00190-009-0308-3.

Mahmood, R., R.A. Pielke Sr., K.G. Hubbard, D. Niyogi, G. Bonan, P. Lawrence, B. Baker, R. McNider, C. McAlpine, A. Etter, S. Gameda, B. Qian, A. Carleton, A. Beltran-Przekurat, T. Chase, A.I. Quintanar, J.O. Adegoke, S. Vezhapparambu, G. Conner, S. Asefi, E. Sertel, D.R. Legates, Y. Wu, R. Hale, O.W. Frauenfeld, A. Watts, M. Shepherd, C. Mitra, V.G. Anantharaj, S. Fall, R. Lund, A. Treviño, P. Blanken, J. Du, H.-I. Chang, R. Leeper, U.S. Nair, S. Dobler, R. Deo, J. Syktus, 2010: Impacts of land use/land cover change on climate and future research priorities. – Bull. Amer. Meteor. Soc. **91**, 37–46, DOI: 10.1175/2009BAMS2769.1.

Meehl, G.A., L. Goddard, J. Murphy, R.J. Stouffer, G. Boer, G. Danabasoglu, K. Dixon, M.A. Giorgetta, A.M. Greene, E. Hawkins, G. Hegerl, D. Karoly, N. Keenlyside, M. Kimoto, B. Kirtman, A. Navarra, R. Pulwarty, D. Smith, D. Stammer, T. Stockdale, 2009: Decadal prediction. can it be skillful?. – Bull. Amer. Meteor. Soc. **90**, 1467–1485.

Müller, W.A., J. Baehr, H. Haak, J.H. Jungclaus, J. Kröger, D. Matei, D. Notz, H. Pohlmann, J.S. von Storch, J. Marotzke, 2012: Forecast skill of multi-year seasonal means in the decadal prediction system of the Max Planck Institute for Meteorology. – Geophys. Res. Lett. **39**, published online, DOI: 10.1029/2012GL053326.

Murphy, A.H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. – Mon. Wea. Rev. **116**, 2417–2424, DOI: 10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2.

Petit, G., B. Luzum, 2010: IERS Conventions (2010). – IERS Technical Note **36**, Bundesamt für Kartographie und Geodäsie, Frankfurt am Main.

Pierce, D., T. Barnett, R. Tokmakian, A. Semtner, M. Maltrud, J. Lysne, A. Craig, 2004: The ACPI Project, Element 1: Initializing a Coupled Climate Model from Observed Conditions. – Climatic Change. **62**, 13–28, DOI: 10.1023/B:CLIM.0000013676.42672.23.

Pohlmann, H., W.A. Müller, K. Kulkarni, M. Kameswarrao, D. Matei, F.S.E. Vamborg, C. Kadow, S. Illing, J. Marotzke, 2013: Improved forecast skill in the tropics in the new MiKlip decadal climate predictions. – Geophys. Res. Lett. **40**, 5798–5802, DOI: 10.1002/2013GL058051.

Reick, C.H., T. Raddatz, V. Brovkin, V. Gayler, 2013: Representation of natural and anthropogenic land cover change in MPI-ESM. – JAMES **5**, 459–482, DOI: 10.1002/jame.20022.

Savcenko, R., W. Bosch, 2012: EOT11a – Empirical Ocean Tide Model from Multi-Mission Satellite Altimetry. – DGFI Report No. **89**, Deutsches Geodätisches Forschungsinstitut (DGFI), München.

Schmidt, T., L. Schoon, H. Dobslaw, K. Matthes, M. Thomas, J. Wickert, 2016: UTLS temperature validation of MPI-ESM decadal hindcast experiments with GPS radio occultations. – Meteorol. Z. **25**, 673–683, DOI: 10.1127/metz/2015/0601.

Seneviratne, S.I., R. Stöckli, 2007: The role of land-atmosphere interactions for climate variability in Europe. – Advances in Global Change Research **33**, Springer Netherlands, 179–193.

Seneviratne, S., P. Viterbo, D. Lüthi, C. Schär, 2004: Inferring changes in terrestrial water storage using ERA-40 reanalysis data: the Mississippi River basin. – J. Climate **17**, 2039–2057, DOI: 10.1175/1520-0442(2004)017<2039:ICITWS>2.0.CO;2.

Seneviratne, S.I., M. Wilhelm, T. Stanelle, B. van den Hurk, S. Hagemann, A. Berg, F. Cheruy, M.E. Higgins, A. Meier, V. Brovkin, M. Claussen, A. Ducharne, J.-L. Dufresne, K.L. Findell, J. Ghattas, D.M. Lawrence, S. Malyshev, M. Rummukainen, B. Smith, 2013: Impact of soil moisture-climate feedbacks on CMIP5 projections: First results from the GLACE-CMIP5 experiment. – Geophys. Res. Lett. **40**, 5212–5217, DOI: 10.1002/grl.50956.

Smith, D.M., S. Cusack, A.W. Colman, C.K. Folland, G.R. Harris, J.M. Murphy, 2007: Improved surface temperature prediction for the coming decade from a global climate model. – Science **317**, 796–799, DOI: 10.1126/science.1139540.

Smith, D.M., R. Eade, N.J. Dunstone, D. Fereday, J.M. Murphy, H. Pohlmann, A.A. Scaife, 2010: Skilful multi-year predictions of Atlantic hurricane frequency. – Nat. Geosci. **3**, 846–849, DOI: 10.1038/ngeo1004.

Smith, D., R. Eade, H. Pohlmann, 2013: A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction. – Climate Dyn. **41**, 3325–3338, DOI: 10.1007/s00382-013-1683-2.

Spangehl, T., M. Schröder, S. Stolzenberger, R. Glowienka-Hense, A. Mazurkiewicz, A. Hense, 2015: Evaluation of the MiKlip decadal prediction system using satellite based cloud products. – Meteorol. Z. **25**, 695–707, DOI: 10.1127/metz/2015/0602.

Stevens, B., M. Giorgetta, M. Esch, T. Mauritsen, T. Crueger, S. Rast, M. Salzmann, H. Schmidt, J. Bader, K. Block, R. Brokopf, I. Fast, S. Kinne, L. Kornblueh, U. Lohmann, R. Pincus, T. Reichler, E. Roeckner, 2013: Atmospheric component of the MPI-M Earth System Model: ECHAM6. – JAMES **5**, 146–172, DOI: 10.1002/jame.20015.

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

Tapley, B.D., S. Bettadpur, M. Watkins, C. Reigber, 2004: The gravity recovery and climate experiment: Mission overview and early results. – Geophys. Res. Lett. **31**, published online, DOI: 10.1029/2004GL019920.

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

Wahr, J., M. Molenaar, F. Bryan, 1998: Time variability of the Earth's gravity field: Hydrological and oceanic effects and their possible detection using GRACE. – J. Geohys. Res. B **103**, 30205–30229, DOI: 10.1029/98JB02844.