npj | climate and atmospheric science

Article

Published in partnership with CECCR at King Abdulaziz University

https://doi.org/10.1038/s41612-024-00788-x

Observations indicate regionally misleading wetting and drying trends in CMIP6

Check for updates

Laura Jensen[1,2] ✉, Helena Gerdener[3], Annette Eicker[2], Jürgen Kusche[3] & Stephanie Fiedler[4,5,6]

We evaluate trends in terrestrial water storage over 1950–2100 in CMIP6 climate models against a new global reanalysis from assimilating GRACE and GRACE-FO satellite observations into a hydrological model. To account for different timescales in our analysis, we select regions in which the influence of interannual variability is relatively small and observed trends are assumed to be representative of the development over longer periods. Our results reveal distinct biases in drying and wetting trends in CMIP6 models for several world regions. Specifically, we see high model consensus for drying in the Amazon, which disagrees with the observed wetting. Other regions show a high consensus of models and observations suggesting qualitatively correctly simulated trends, e.g., for the Mediterranean and parts of Central Africa. A high model agreement might therefore falsely indicate a robust trend in water storage if it is not assessed in light of the observed developments. This underlines the potential use of maintaining an adequate observational capacity of water storage for climate change assessments.

Global warming and the intensification of the water cycle are expected to cause persistent drying and wetting patterns[1,2], with consequences for water resources, food, health, transport, tourism, and other sectors. Physical arguments entail that the precipitation − evapotranspiration deficit $P - E$ will be enhanced, owing to an increased water-holding capacity of the atmosphere. This explains only a wet gets wetter (WGW) response over land since, assuming equilibrium state, the discharge Q cannot exceed $P - E$. However, WGW has been observed for a fraction (10.8%) of the global land only[3]. Simulations suggest both WGW and dry gets drier (DGD) emerging over land[1] and oceans[4] at multidecadal timescales, inspiring the term WGWDGD, although recent observational studies disagree[3,5–7].

Global coupled Earth System Models (ESMs) are used for simulating the future emergence or amplification of wetting and drying under socio-economic scenarios, and for attributing to potential drivers. It is thus relevant in how far such patterns can be confirmed with observations. Wetting and drying can be understood as trends in the mean or occurrence of more frequent and heavier extremes, and both must be considered since e.g., DGD implies higher-impact droughts. However, constructing observational records is challenging due to heterogeneity of the observing system in terms of coverage and sampling, and the stochastic nature of the climate system. The joint presence of trends, variability, and extremes causes difficulties in observational studies with limited temporal coverage[6].

Another question is what observable one should consider for quantifying wetting and drying[8]. This was originally studied in terms of $P - E$, while recently other metrics have evolved[8]. In summary, such studies (based on models or observations directly) showed robust WGWDGD patterns for few areas only, whereas overall a more complex picture of wetting and drying than previously suggested[1] was found[3,5,8,9]. These studies assume that change in water storage dS/dt is negligible, i.e., $P - E$ balanced by Q. However, since the Gravity Recovery and Climate Experiment (GRACE)[10] satellites began measuring in 2002, we know that total water storage changes (dS/dt) occur at all spatial and temporal scales, within the duration of the data record at least[11]. Runoff and discharge depend on land surface and may not balance $P - E$; as a result, $P - E - Q$ represents an indicator of land water storage changes due to natural droughts or pluvials as well as direct and indirect anthropogenic actions[12], and intensification causes departure from the equilibrium $P - E - Q = 0$ even at the global scale[13,14]. Considering only the departure, including activities like groundwater pumping, irrigation, and reservoir management (I), the effects of which are actually observable with GRACE[15–18], the evolving balance can be viewed as $\Delta P - \Delta E - \Delta Q - I = dS/dt$. This means TWSA is linked to (i) modifications of land conditions and resulting climate forcing ($\Delta P - \Delta E$), (ii) direct and indirect human impacts ($I + \Delta E$), and (iii) hydrological response of the system ($\Delta Q + \Delta E$)[6,19,20]. Along this line GRACE TWSA trends can be

[1]Helmholtz Centre Potsdam - GFZ German Research Centre for Geosciences, Potsdam, Germany. [2]Geodesy and Geoinformatics, HafenCity University, Hamburg, Germany. [3] Institute of Geodesy and Geoinformation, University of Bonn, Bonn, Germany. [4]GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany. [5]Faculty of Mathematics and Natural Sciences, Kiel University, Kiel, Germany. [6]Present address: Institute of Environmental Physics, University of Heidelberg, Heidelberg, Germany. ✉e-mail: laura.jensen@gfz-potsdam.de

attributed to 'climate change impact', 'direct human impact', and 'natural variability'[21].

TWSA as measured by GRACE refer to the entire land system including changes in all water storage compartments (soil moisture, surface water, snow, groundwater). Long-term changes had been difficult to assess from modeling[22], but we now have twenty years observations from GRACE and its follower GRACE-FO[23–25]. Previous studies compared GRACE trends to atmospheric reanalyses or models forced with reanalyses, finding disagreement at interannual timescales, and suggesting opposite behavior (water net gain or loss) in major catchments compared to GRACE[6,22]. In a comparison of GRACE and discharge data to reanalyses, $P-E$ offsets were attributed to precipitation biases[26]. Recently, consistent biases in seasonal TWSA derived from ESM ensembles compared to GRACE were found, suggesting that many models overestimate precipitation, with incorrect representation of runoff as another contributing factor[27,28]. Another study[7] compares GRACE reconstructions, hydrological models, and ESMs, quantifying drying/wetting via trends in storage-based hydrological droughts[29,30].

On the downside, the GRACE record has gaps due to instrument problems and since GRACE-FO could not be launched in time for continuation. The data have an effective resolution of 300 km and this is not expected to significantly change even beyond GRACE-FO[31]. Also, the data contain geophysical signals not related to water masses (e.g., earthquakes and glacial isostatic adjustment (GIA))[24] as well as anthropogenic impacts (e.g., groundwater withdrawal) not considered in climate models. They also include reservoir storage changes which are not well represented in hydrological models, but for large reservoirs this can be accounted for[32].

State-of-the-art ESMs, for example those taking part in the Coupled Model Intercomparison Project Phase 6 (CMIP6)[33], are indispensable tools for gaining future climate information but are of course far from perfect. Besides simplified process representations[34,35], mainly inaccurate parameterization has been shown to be a reason for limitations and uncertainties in ESMs, which can result in misleading conclusions about feedback cycles[36,37]. Also soil moisture, the main water storage-related variable in ESMs, is highly dependent on complex soil and vegetation characteristics, which are often not well-known on a global scale. Simplified assumptions on processes and parameters affecting soil moisture estimates may also feedback into land-atmosphere interactions, further amplifying uncertainties[38–40].

In a previous study[41] we derived hot spots of wetting and drying in land water storage from a CMIP5[42] ensemble and a 14-year GRACE record, while masking regions where geophysical or anthropogenic processes could not be ruled out. High-consensus regions of 250-year (1850–2100) modeled total water storage (mTWS) trends were identified. However, fitted linear trends in the short observational record were dominated by interannual variability, and we suggest this to cause mismatch between GRACE and model trends. In fact, this study revealed that the chance of a match between 14- and 250-year trend is only 53% when averaged over the globe. However, we emphasize at regional scales there are regions where variability is less pronounced and the sign of the trend much less sensitive with respect to the time series length.

Given these limitations, we revisit whether modern climate models reproduce wetting/drying patterns from observations. We use CMIP6 data and extended observation data until 2019/12 (instead of 2016/08) to perform a more comprehensive identification of possible long-term climate trends in land water storage. In addition to identifying regions of large inter-model consensus, we determine regions of temporal consensus between trends computed over different periods, i.e., regions with temporally consistent trends over longer time periods. This step is motivated by the natural variability in water storage that is not forced by anthropogenic perturbations. We assess where long-term model trends can be expected to meet (shorter) observational trends. Further, we partition the CMIP6 ensemble into models with deep and shallow soil layers. Instead of using original GRACE/-FO observations, we make use of a new reanalysis product for Global Land Water Storage (GLWS) data[43] which was derived via assimilation[44] of GRACE and GRACE-FO TWSA data into the offline WaterGAP[45] hydrology model. GLWS provides monthly TWSA data that is spatio-temporal consistent and at a higher spatial resolution compared to GRACE/-FO. It does not suffer from data gaps, and is less reliant on the GRACE instrument data quality. More realistic TWS trends can be derived from GLWS, uncertainty estimates are more comprehensive as compared to earlier assimilations[43,46], and this will enable better comparisons to climate model simulations. Furthermore, GLWS provides estimates for the individual storages, namely soil moisture, surface water, snow, and groundwater, which the satellite product alone does not separate, and which may pave the way for more comprehensive investigations of climate model deficiencies.

Results

Consensus of mTWS trends in CMIP6 model output

We investigate (1) the consensus among individual models of the CMIP6 ensemble, regarding the direction (sign) of long-term trends and (2) the agreement of short-term with long-term trends. The rescaled multi-model median (MMMed) mTWS trend from 17 CMIP6 ESMs over 1950–2100 is displayed in Fig. 1a, cf. Methods section. mTWS trends of up to 4.3 mm/yr and down to −23.6 mm/yr occur globally, with drying prevailing (58.6% of the land area excluding Greenland, Svalbard, and Antarctica) over wetting trends (41.4%). The model spread of trends is quite large (Fig. 1b). Therefore, to identify regions with good agreement among models we derive the model consensus (see "Methods"), i.e., the number of models agreeing on the trend direction (Fig. 1c) for each grid cell. We restrict further analyses to high-consensus regions, i.e., where at least 75% of models agree on the sign, which cover 48.0% of the land (red and blue color in Fig. 1c). In the remaining 52.0% there is strong disagreement among the models on trend direction (gray color in Fig. 1c). High consensus on long-term drying can, e.g., be identified in the Amazon, around the Mediterranean Sea, the SW U.S., and southern Africa. Models agree on long-term wetting particularly in central Africa and the Sahel, India, and parts of Central Asia.

We investigate the agreement of short-term trends for various time spans with the long-term (150-years) trend and derive regions that represent high temporal consensus (see "Methods"). As expected, with increasing short-term period the average agreement becomes higher: Fig. 2a–c shows the percentage of trend estimates from moving 17-, 30-, and 50-years slices that match the 150-year trend. However, there appears also a distinct spatial pattern of less or more than average agreement; quite robust among different time periods. Grid cells with higher-than-average agreement with the 150-year trend for all periods (17, 20, 25,..., 50 years) appear red in Fig. 2d: these are regions of CMIP6 high temporal consensus mTWS trends (27.6% of the land), which suggests the influence of interannual variations is low. We therefore assume that in these regions we can safely compare long-term model trends with observations. High temporal consensus often corresponds to high model consensus (cf. Fig. 2d and Fig. 1c).

Observation-based trends from GRACE/-FO and GRACE/-FO data assimilation

We estimate trends for both GRACE/-FO and GLWS[43] (referred to as GRACE/DA in the following) data (see "Methods"). Figure 3a shows linear TWS trends for 2003–2019 in the original GRACE/-FO data, with large negative trends with −15 mm/yr or more, e.g., in East Brazil, Caspian Sea, parts of India and glacier-rich Alaska, Patagonia, and Canada Northern territories. In previous studies[21,47–51], these regions were identified to match with droughts, decreasing water surface levels, groundwater depletion, and glacier melting. Large positive trends can be found in Central Canada, West and Central Africa, and in South Brazil/North Argentina, which were all identified as affected by increasing precipitation, e.g., recovery after a dry period at the beginning of the observation period[21].

TWS trends for the same period (2003–2019) but from GRACE/DA are shown in Fig. 3b. Regions of pronounced drying and wetting coincide with the original GRACE/-FO data, with the main difference being spatial resolution. For example, large negative GRACE/-FO trends in Patagonia appear more focused on actual glacier locations in GRACE/DA. Due to the higher spatial resolution, we use GRACE/DA in what follows for comparison to CMIP6. However, we aggregate GRACE/DA and ESM output to 2° since the resolution of most ESMs is currently not finer.

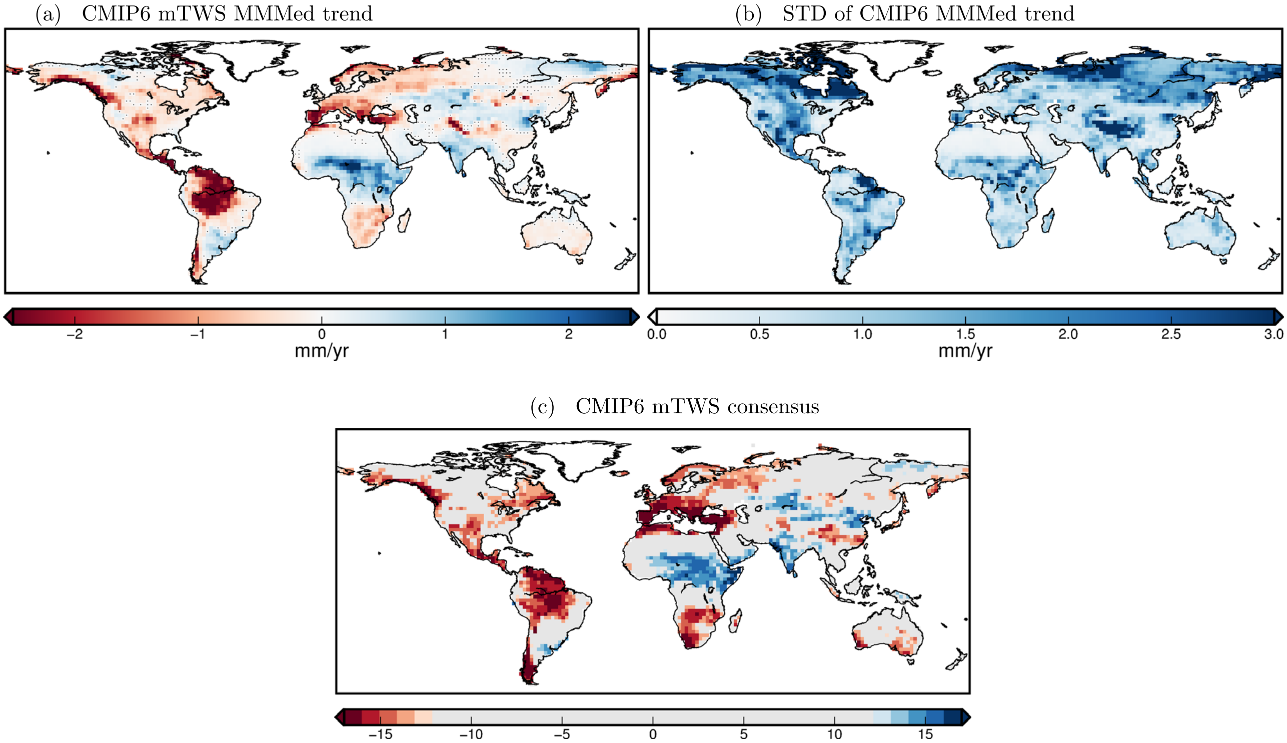

Fig. 1 | Analysis of mTWS trends in the CMIP6 ensemble. a Rescaled weighted multi-model median (MMMed) trend (1950–2100) in each grid cell, stippling indicates non-significant trends; **b** standard deviation of trend values; **c** the model consensus showing the number of models agreeing on the direction of the drying (red) and wetting (blue) trend.

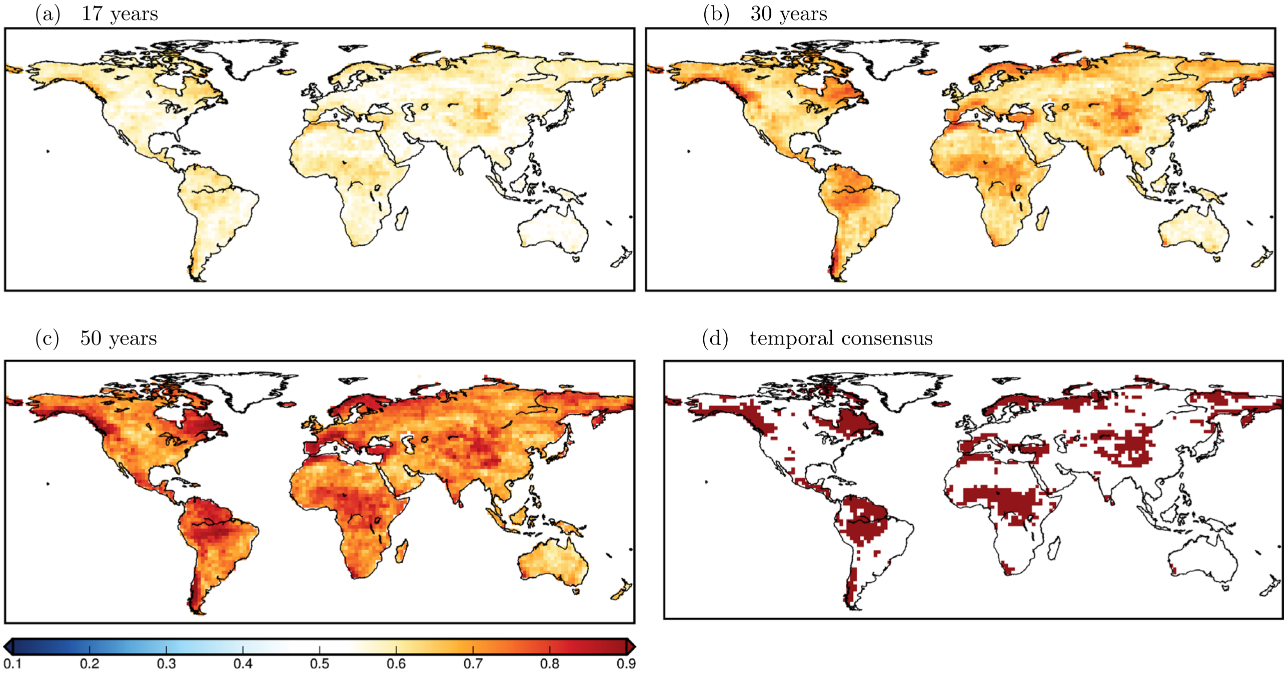

Fig. 2 | Temporal consensus. a–c Agreement of trend estimates derived from (overlapping) short-term time slices (derived from 17, 30, and 50 years of model output) with the 150-years model trend. **d** Grid cells with higher-than-average agreement with the 150-year trend for all periods (17, 20, 25, 30, 35, 40, 45, 50 years).

Agreement of high consensus mTWS trends with assimilated TWS trends

Observed GRACE/DA trends (Fig. 3b) are compared to CMIP6 MMMed mTWS trend estimates. We focus on regions in which the model ensemble shows strong agreement in trend sign and estimates appear robust over different time spans. For this purpose, Fig. 4a merges regions of (1) high model consensus (Fig. 1c) with regions having (2) a high temporal consensus (Fig. 2d). All colored grid cells in Fig. 4a exhibit high model consensus and cells lacking high temporal consensus are shaded in gray. The map is then compared to the sign of the observed trends (Fig. 3b), with dark colors indicating agreement of modeled and observed trend direction and light colors for disagreement.

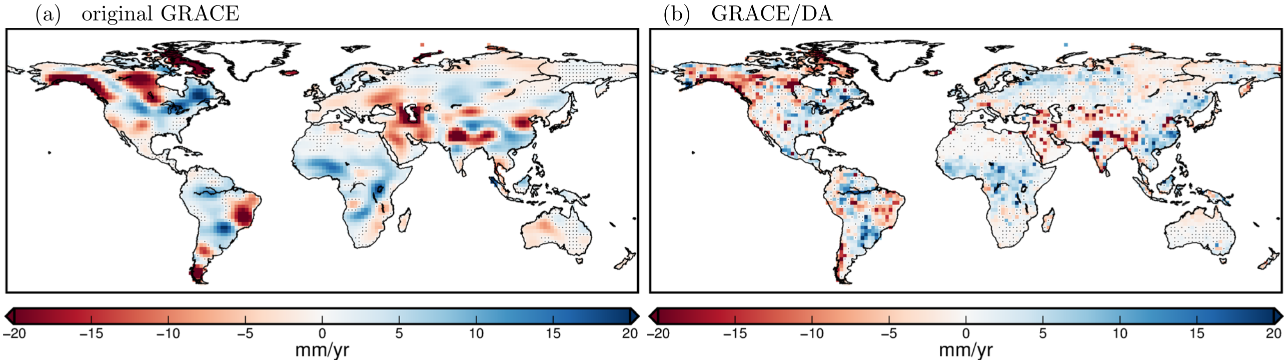

Fig. 3 | Observation-based linear TWS trends. Trends computed over 2003–2019 for **a** the original GRACE/-FO data that were used as input for the data assimilation, **b** the GRACE/DA product.

For clarity, in Fig. 4b–e we plot different parts of Fig. 4a, regarding different combinations of model and temporal consensus and their consistency with observational trends. We introduce the notation $M^{+/-}T^{+/-}O^{+/-}$, where M stands for model consensus, T for temporal consensus, and O for observations, with high (+) or low (−) consensus, and identical (+) or opposite (−) sign of observational trend. For example, $M^+T^+O^+$ denotes regions in which models exhibit high model (M^+) and high temporal consensus (T^+) and agree with observations (O^+) (Fig. 4b). Complementary to regions of high model consensus (M^+, Fig. 4a, b–e) are regions of low model consensus (M^-, Fig. 4f–i). In regions with low model and low temporal consensus (M^-T^-, Fig. 4h and i), the fit to observations is irrelevant as there appears no distinct model agreement for any time period. In the subsequent section, we discuss selected regions from Fig. 4b to i with respect to possible causes for (dis)agreement.

Globally, high model consensus areas (M^+) make up about 48.0% of the land area, and in 44.5% of high model-consensus areas also the temporal consensus is high (M^+T^+). In M^+T^+ regions, the model trend is confirmed by GRACE/DA trends in about 63.0% of the area ($M^+T^+O^+$). In contrast, in 37.0% of the land the GRACE/DA trend shows the opposite sign ($M^+T^+O^-$). When temporal consensus is disregarded, the area of agreement between model and GRACE/DA trend is only about 55.2% ($M^+T^{+/-}O^+$), which stresses the importance of consistency of trends over different time periods.

Disagreement between the majority of ESMs and observations may be caused by limitations in soil parameterization, resulting in deficiencies in evaporation and soil–groundwater interactions[52,53]. One parameter influencing simulated soil moisture is the total soil depth of the ESMs. We partition the CMIP6 ensemble into two subsets, one containing all ESMs utilizing a land surface model with a thickness of all soil layers of >4.6 m (deep soil), and one with ≤4.6 m (shallow soil). We repeat the trend computation and derive the overlay map of model/temporal consensus and observational trend for the subsets (Fig. 5). Trend maps (Fig. 5a, b) are similar regarding the spatial pattern but the magnitude of the trends is larger for the deep soil model ensemble in many regions. We explain this by a larger water-holding capacity in models with deep soil, thus increasing storage times and enhancing the simulation of long-term trends. The M^+ region is smaller for deep soil than for shallow soil models (45.8% vs. 53.9%, Fig. 5c, d), which may indicate a larger complexity (and thus discrepancy) for models with deep soil. The extent of model-observation agreement ($M^+T^+O^+$) is very similar for both model subsets (62.1% vs. 61.4%). We conclude that soil moisture representation in ESMs is complex[54], with soil depth being only one contributing factor, and that a deep soil parameterization may be more advantageous for simulating realistic intra- and interannual storage variability than for long-term trends. However, the two subsets of models are rather small (9 ESMs with shallow and 8 ESMs with deep soil), which somewhat limits the robustness of this comparison. Furthermore, the threshold of 4.6 m (median soil depth of all models) for the partitioning is not physically based.

Analysis of selected regions

The comparison of TWSA sign from CMIP6 ESMs and GRACE/DA (Fig. 4a) reveals distinct geographical regions, some of which are discussed in the following because they either (1) show good agreement between models and observation in high consensus regions ($M^+T^+O^+$), (2) exhibit distinct disagreement despite high model consensus ($M^+T^+O^-$), or (3) show clear mTWS trend, but low consensus either across models and/or regarding temporal consistency of modeled trends ($M^+T^-O^+$, $M^-T^+O^+$, or $M^-T^-O^{+/-}$). In the upper panel of Fig. 6 regions with different +/− combinations of M-T-O are outlined. In the lower panel, the mTWS trends for all 17 ESMs are shown spatially averaged over each region. The assignment of the trend values to the regions is possible via the colors at the top of the lower panel.

High model and temporal consensus with GRACE/DA agreement ($M^+T^+O^+$).

In $M^+T^+O^+$ regions (Fig. 4b), a majority of ESMs agree on wetting/drying trends, short-term trends are supposed to be representative of the long-term trend, and observations support the direction of ESM trends. Thus, we suggest these regions are suspect to climate-related wetting or drying. However, agreement between model and observational trends still has to be interpreted with care, as there are processes not included in ESMs but present in observations. For example, the negative mass trend in Alaska is mainly due to glacier melting, which is not yet considered in ESMs, rendering a comparison in this region difficult as long as this process cannot be properly separated from the observational record.

Mediterranean region and Turkey. ESMs exhibit drying with very strong model consensus (region average 97.3%), and also high temporal consensus. There is a relatively small model spread in the mean trend (Fig. 6), the ensemble mean trend is negative for all models, and drying is also prevalent in GRACE/DA. This region was already identified as a hot spot in CMIP5[41]. Increasing drying conditions are confirmed by several other studies. For example, in the historical period (1948–2005), Greve et al.[3] found South Europe as WGD, except Spain as DGD, but no significant change in Turkey. In a follow-up study, Greve et al.[8] suggest South Europe including Turkey as WGD and TGD (transitional, no significant attribution of current state), under RCP8.5 and 2100–2080 vs. 2000–1980. The driver of the drying is probably a combination of decreasing precipitation (in the western) and increasing evapotranspiration due to increasing temperatures (in the eastern part), which is seen for 1980–2014 in observations from the Climatic Research Unit (CRU)[55], resulting in $P-E$ being significantly negative[8]. As a caveat, several dams have been built in the *Güneydoğu Anadolu Projesi* (GAP) since about 1990; this includes the large Ataturk and Karakaya reservoirs and Turkey's biggest lake. Abdelmohsen et al.[18] show how GAP river regulation effectively buffers droughts and floodings at a scale visible from GRACE/-FO. Therefore, we cannot expect a perfect agreement with models.

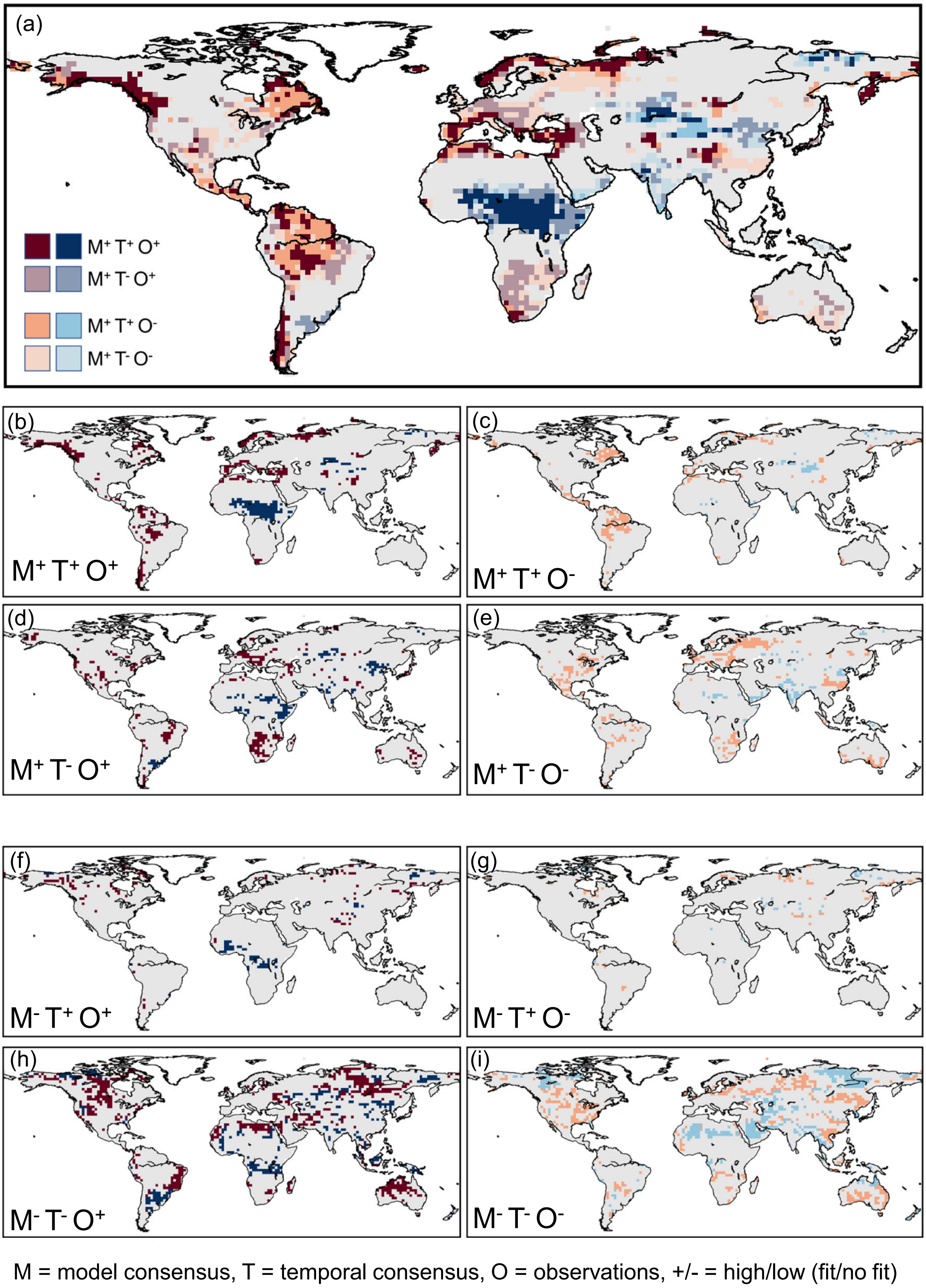

Fig. 4 | Agreement of high consensus CMIP6 mTWS trends with observation-based trends. a Regions of high model consensus of 17 CMIP6 models regarding the direction of mTWS trends, distinguished into high/low temporal consensus and fit/no fit to GRACE/DA TWS trends. See text for explanation of legend notation. **b**–**e** Different subsets of (**a**), fulfilling different criteria regarding high/low temporal consensus and fit/no fit to observations. **f**–**i** Regions of low model consensus, and different combinations of high/low temporal consensus and fit/no fit to observations.

Central Africa. In this vast region, including Congo and Lake Victoria basins, parts of the Niger, and the White Nile, ESMs exhibit wetting with strong model (region average 85.9%) and high temporal consensus, and this is also projected in the near-term under all SSPs[56]. However, there appears relatively large model spread (Fig. 6). The wetting trend is supported by GRACE/DA. This region was not yet identified as a hot spot in CMIP5[41] due to less model consensus and insufficient support by GRACE/-FO observations. Pronounced overall wetting in GRACE/DA can be traced to changes in

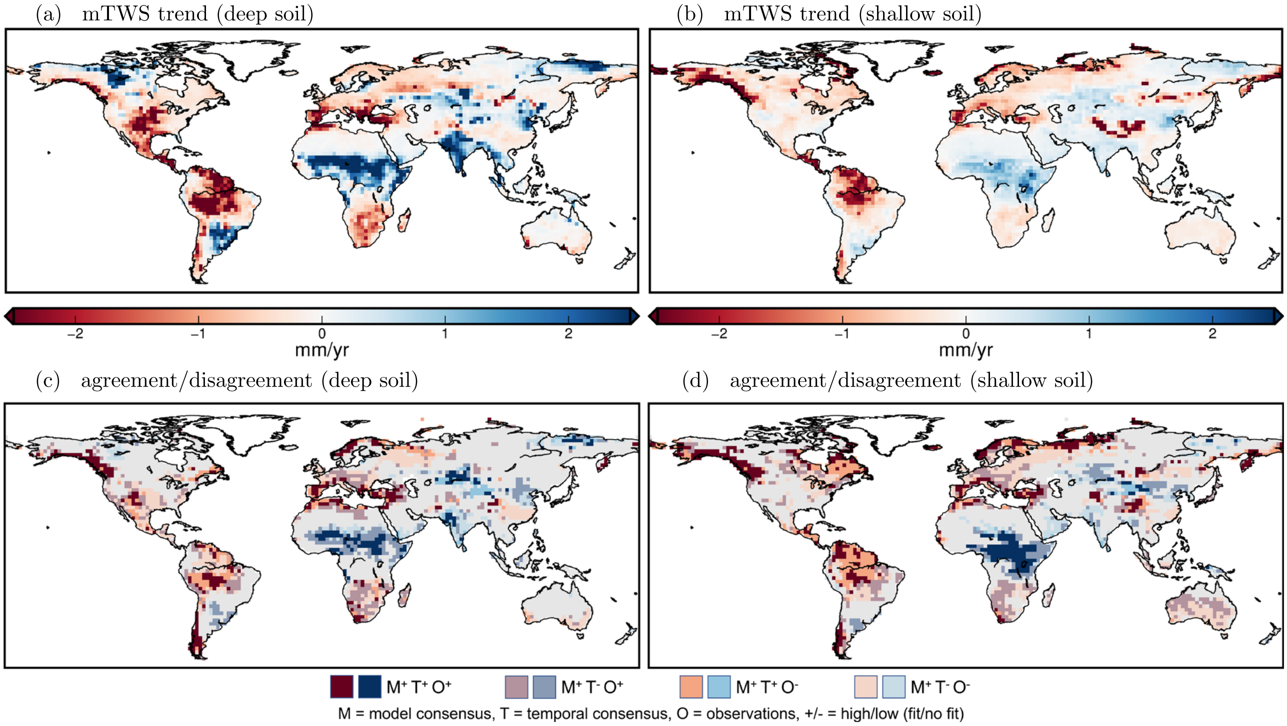
(a) mTWS trend (deep soil)
(b) mTWS trend (shallow soil)
−2
−1
0
1
2
mm/yr
−2
−1
0
1
2
mm/yr
(c) agreement/disagreement (deep soil)
(d) agreement/disagreement (shallow soil)
$M^+T^+O^+$
$M^+T^-O^+$
$M^+T^+O^-$
$M^+T^-O^-$
M = model consensus, T = temporal consensus, O = observations, +/- = high/low (fit/no fit)

Fig. 5 | Partitioning the model ensemble into deep and shallow soil ESMs. Above: Rescaled weighted mTWS MMMed trend (1950–2100) for models with a total thickness of all soil layers of **a** >4.6 m (deep soil; 8 models), and **b** ≤4.6 m (shallow soil; 9 models). Below: Regions of high model consensus regarding the direction of mTWS trends, distinguished into high/low temporal consensus and fit/no fit to GRACE/DA TWS trends for **c** deep soil models, and **d** shallow soil models.

precipitation regimes since 2015. Recent assessments therefore conflict with earlier studies, which found Central Africa as WGD[3] and suggested that $P - E$ changes play only a minor role and WGWDGD thus being insignificant in this region[8].

Congo has a bimodal rainfall distribution with large year-to-year variability, precipitation and water storage are difficult to simulate due to large land surface feedbacks, deep soils[57] poorly represented in models, and few measurements allowing model calibration. The hydrological model that underlies GRACE/DA exhibits little interannual variations and overall wetting/drying of a few mm/yr only. While Rodell et al.[21] ascribe decreasing storage in GRACE data 2002–2016 to precipitation decrease, this reversed since then, with record precipitation in 2019. In GRACE/DA this is attributed to a sudden increase in soil wetness since 2018, only partly compensated by ongoing groundwater decrease. In addition, a sustained trend of several cm/yr in basin-averaged surface storage accumulation since 2006 contributes to the wetting trend, whereas groundwater storage in GRACE/DA shows little net effect.

The Niger region is characterized by unimodal rainfall distribution, with the inner Niger Delta turning into a floodplain every year, another modeling challenge. GRACE, GRACE/DA, and the underlying hydrological model consistently suggest wetting of several cm/yr, mainly due to sustained recovery prior to 2012 in both groundwater and surface storages.

Observed TWSA in the White Nile and Lake Victoria basins are attributable to reservoir construction and managed lake levels, as several authors demonstrated. In fact, GRACE/DA reveals significant (few cm basin-averaged for Lake Victoria) interannual variability in surface and groundwater storages, but little trend over the entire time frame. These effects explain a misfit of observations to models, but when averaged over large regions they are dwarfed by the large precipitation signal in the Congo.

High model and temporal consensus with GRACE/DA disagreement ($M^+T^+O^-$). In $M^+T^+O^-$ regions (Fig. 4c) a majority of ESMs agree on trend direction, and short-term trends are consistent with centennial trends, but observations have the opposite direction. The mismatch can either hint to ESM deficiencies or to remaining influence of interannual variations. Such regions are particularly interesting because high CMIP6 model consensus with low spread suggests confidence, but this can be misleading and poses the risk of inadequate measures for climate change action.

Amazon/Orinoco region. Amazon and Orinoco show particularly strong CMIP6 model consensus (region average 93.4%) on drying, with low spread (Fig. 6), but in disagreement with GRACE/DA. Models also exhibit a mean underestimation of precipitation[34]. Again, model precipitation biases are pronounced during monsoon season and persist since CMIP3, but reduced in recent CMIP phases[34]. Similarly, regional drying was seen across CMIP5, but wetting in GRACE[41]. In Jensen et al.[41] recovery from drought in the early GRACE period was suggested as a possible reason for the wetting trend. While such multiannual variations may still overlay long-term observational trends, their influence is decreasing with increasing duration, and there are also indications for a WGW regime in this area, even though rather localized, and possibly not statistically significant yet[3,8].

Mismatch can result from systematic model deficiencies. The drying trend in CMIP6 mTWS is most likely driven by a negative trend in precipitation, but CMIP6 underestimates precipitation across the Amazon measured against observational products (CRU, TRMM), which is at least in parts associated with precipitation biases during El Niño events with little improvements across CMIP phases[34]. In addition, missing groundwater in CMIP6 models may result in artificial drying trends. It was shown that considering groundwater buffering in CMIP5 models causes a shift in the evapotranspiration regime resulting in less drying trends in the Amazon[58]. Soil depth in this region is heterogeneous, but mostly deep[57]. The hypothesis that model deficits cause drying is further supported when considering only deep-soil models. For models with deeper soils, the area of high temporal consensus in the Amazon and Orinoco region is much smaller compared to the entire CMIP6 models ensemble (Fig. 5c), which means that some models with a deep soil indeed simulate the observed wetting at least over shorter time periods.

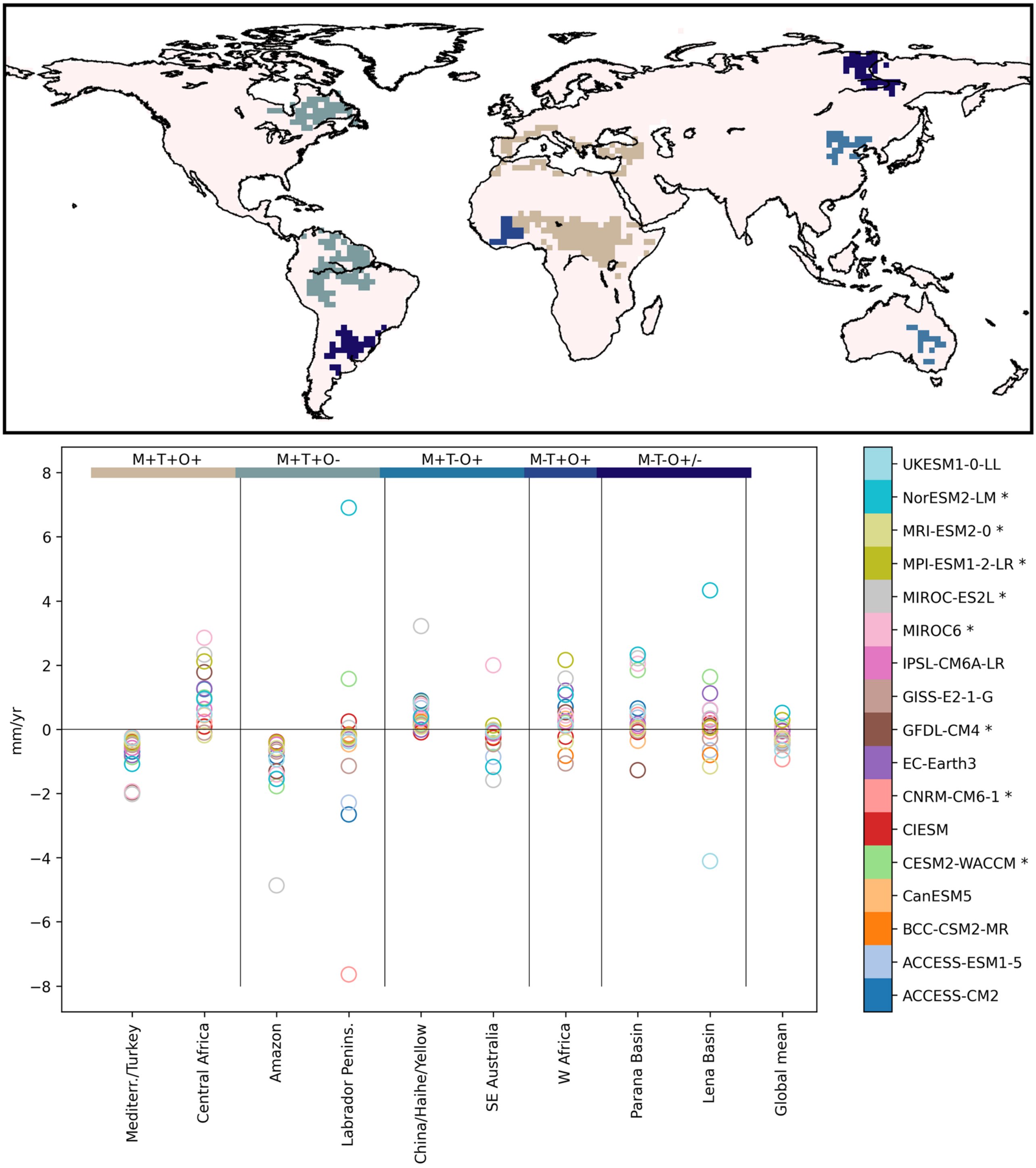

Fig. 6 | Spatial mean mTWS trend over selected regions for 17 CMIP6 models. Different colors in the map denote different cases of model/temporal consensus and fit/no fit to observations, as indicated in the lower panel. Models marked with an * in the legend denote models with deep soil (>4.6 m) in the land surface model.

Labrador Peninsula. Here the average model consensus is high with 83.5% agreement on drying, yet with large spread (Fig. 6) and in disagreement with GRACE/DA. In CMIP5 this was also a high consensus region[41]. A reason for wetting trend may be residual GIA trends in GRACE/DA; it was shown recently that in this region the choice of the GIA correction can significantly affect the observational trend[59]. However, despite this, this region is assumed to exhibit WGW regime with $P-E$ being significantly positive[8]. Regarding the model trends, no clear precipitation bias for CMIP6 is identified for the Labrador Peninsula relative to CRU, but there is observational evidence for CMIP6 models wrongly simulating decreasing snow depth trends[60]. Furthermore, when considering only CMIP6 models with deep soil, there is no agreement on drying conditions in the northern part of the Peninsula (Fig. 5c), which means that the observed wetting can be reproduced by some deep soil models and thus strengthens the hypothesis that deep-soil representation aids in simulation of TWS dynamics.

High model but low temporal consensus and GRACE/DA agreement ($M^+T^-O^+$).

In $M^+T^-O^+$ regions (Fig. 4d) a majority of ESMs agree on trend direction, but low temporal consensus indicates that the observations (agreeing with the direction of the models) may not be representative for the long-term trend. In such regions, interannual variations overlaying long-term climate signals may play a particularly strong role and agreement of models and GRACE/DA might only be a coincidence.

Haihe/Yellow river region (China). Models show an average consensus of 83.0% on wetting trend, with small spread (with one exception), and GRACE/DA also hints at wetting. However, there is no high temporal consensus, which suggests that it might be difficult to compare long-term trends and observations. In this region the advantage of the increased resolution of GRACE/DA becomes obvious: due to the spatial filtering of the original observations a strong drying in the Haihe/Yellow, probably due to groundwater use[21], is blurred. GRACE/DA reveals a differentiated pattern with drying and wetting, and large-scale coherent trend in the Haihe/Yellow river, which is also seen in ESMs. Wetting is supported by a small positive bias in $P-E$[8], even though it is not significant.

Southeast Australia. Models agree on drying with 79.5% average model consensus and small spread (with one exception). While GRACE/DA agrees, it is not clear to what extent short-term trends can be compared to the long-term trend, as there is limited temporal consensus among models. In SE Australia interannual variations are significant: it experienced a very severe *Millenium Drought* in 2001–2009[61], and 20-year trends may be masked by the recovery. For example, Rodell et al.[21] find the region in GRACE data as slightly wetting and ascribe this to progression from dry to wet conditions; however, their time frame is 2002–2016, and 2017–2019 were characterized by unusual droughts, leading to an overall drying trend. Also, surface-water groundwater interactions seem to play a significant role, which even hydrological models are not able to correctly reproduce[62]. Therefore, it is unclear whether the agreement is more than a coincidence.

Low model but high temporal consensus and GRACE/DA agreement ($M^-T^+O^+$). While model consensus is relatively low in $M^-T^+O^+$ (Fig. 4f), the observational trend is likely representative for the long-term trend. This is an indication that the MMMed mTWS trend is real, even though it is currently only supported by a small majority of models.

West Africa. ESMs show wetting with average model consensus of only 69.1%, and large spread, but high temporal consensus. Also, GRACE/DA suggests a wetting trend here. A reason for low model consensus may be that land-atmosphere coupling plays a large role[54] and feedbacks are important that almost certainly are not well represented in CMIP models. Rainfall is related to the location of the Intertropical Convergence Zone (ITCZ), which may be positioned differently from model to model, explaining the large model spread. Again, tropical soils are likely poorly represented in models[53]. There is a positive mean precipitation bias in coastal West Africa during summer monsoon in CMIP6 w.r.t. TRMM, which was also simulated in earlier CMIP phases[34], and rising TWS trends are attributed mainly to a precipitation increase[21]. However, also dam building[63], land use change and cropland expansion may contribute to increased storage[64].

Low model and low temporal consensus ($M^-T^-O^{+/-}$). The reported MMMed mTWS trend in $M^-T^-O^{+/-}$ regions (Fig. 4h,i) is neither supported by high model nor high temporal consensus. In this case, it is reasonable to take the magnitude of the model trend into account. Model disagreement is generally more likely the closer the trend is to zero[41]. Therefore it is interesting to investigate regions with large trend signal (below the 25%- or above the 75%-percentile) despite low consensus.

Parana basin. Models show an average consensus of only 65.4% with preference on wetting, in agreement with GRACE/DA. However, due to low model and temporal consensus it is unclear if this represents a climate signal or rather originates from interannual variability or human interactions. Wetting in observations may be due to recovery from an early-period drought (2001–2005), similar as in the largest part of the Amazon[21]. This drought has been so strong that it has been even visible in TWSA observations from satellite laser ranging[65]. Also the filling of reservoirs in southern Brazil may have contributed to the observed trend[21]. A small (insignificant) positive bias in $P-E$[8] may be driving the wetting in the majority of ESMs. Low model consensus may be due to the varying skill of CMIP6 models in the spatio-temporal distribution of precipitation and temperature found for South America[66].

Lena basin. Wetting with an average model consensus of only 64.4% is reported by ESMs, which contradicts the GRACE/DA trend. The spread is large, with two contradicting extreme trends of −4.1 mm (UKESM1-0-LL) and +4.3 mm (NorESM2-LM). A reason for large model spread and low consensus may be that the basin is governed by permafrost soils, which are challenging to simulate[53]. Furthermore, interannual variations likely conceal long-term trends. For example, a strong TWSA increase in 2001 and 2002 had been reported for the Ob/Yenisei/Lena, most likely a rebound following a steady decline in water storage since 1994[65].

Discussion

From our analysis of inter-model, temporal, and observational consensus, a robust picture of CMIP6 model limitations in simulating drying and wetting in terms of water storage emerges, despite the still short GRACE/-FO record.

Good agreement between models and observations in high consensus regions ($M^+T^+O^+$), is found in ~13% of the global land. For the Mediterranean region and Central Africa, we suggest that drying and wetting are predominantly driven by precipitation, responding to climate change, and that confounding effects in observations such as lake level changes average out. However, due to the superposition of various signals, some of which not taken into account in the same way in observations and models (e.g., glacier melting), in some $M^+T^+O^+$ regions attributing drying/wetting to climate is still challenging.

Regions with distinct disagreement despite high model consensus ($M^+T^+O^-$) cover ~8% of the global land. For Amazon/Orinoco and Labrador Peninsula we argue that either remaining multiannual variations in observations or deficiencies in process simulation of e.g., groundwater-surface interactions or snow accumulation (in either CMIP6 models or the hydrological model behind GRACE/DA) cause the mismatch. These regions are interesting because, despite a strong model consensus, CMIP6 trends can be misleading. We add that in some regions, e.g., the Haihe/Yellow river and South East Australia, model and observational trends agree, however, the temporal consensus of the models is low ($M^+T^-O^+$). These regions extend to ~13% of the global land.

Low model consensus but agreement with observations, which (due to high temporal consensus) seem representative for long-term trends ($M^-T^+O^+$) is found in ~4% of the global land. Complex soil and/or atmospheric conditions, as e.g., in West Africa, may be challenging to model, explaining the low consensus. About 46% of the land surface exhibits low consensus across models and low temporal consistency of modeled trends ($M^-T^-O^{+/-}$). In many regions mTWS trends disappear, but some other regions show clear trends despite low consensus, e.g., Parana and Lena basins. Again, atmospheric or permafrost conditions may explain low model consensus, and strong interannual variations may prevent temporal consensus.

Partitioning the ensemble into subsets with deep (>4.6 m) vs. shallow soil (≤4.6 m) provides only limited insights into model performance. Weighted mTWS trends appear significantly larger for models with deep soil than with shallow soil, with the deep-soil ensemble showing a smaller region of model consensus. While the overall advantage of a deep soil parameterization may be more noticeable for short-term TWS variability rather than long-term trends, there are some hints in model-observation mismatch regions that deep soil models are better capable to meet observational trends. However, with only 9 and 8 models in the subsets these results may not be representative.

Overall, by evaluating mTWS trends from a CMIP6 model ensemble by means of a new GRACE assimilation product (GRACE/DA) we could show that, while in some regions wetting and drying can be confirmed, in several world regions CMIP6 model agreement about trends in water storage is not necessarily indicating consistency with trends from observational data. Using an assimilation product for TWS enabled us to compare to CMIP6 at a spatial resolution of 200 km. We acknowledge that any assimilation gives rise to data depending on a specific model, but we believe this is outweighed by the ability to downscale the GRACE data and the ability to look at individual storages represented in GRACE/DA, i.e., snow water equivalent or soil moisture. We account for uncertainties in modeled TWSA due to structural model deficiencies in representation of hydrological process and land-atmosphere interactions by (1) computing the multi-model median of all available models, (2) considering only regions of large inter-model consensus, and (3) focusing only on the sign of the trend rather than the actual magnitude, assuming that CMIP6 TWS trends mostly reflect wetting/drying of the full water column even though they do not explicitly contain groundwater storage. Furthermore, discrepancies between modeled and observed TWSA are due to CMIP6 neglecting surface waters and glaciers. However, their influence is local and we focused on regions not (or only marginally) affected.

The length of the GRACE/-FO record is steadily increasing. As the essential climate variable TWS, at least in regions unaffected by direct water use, provides inference on precipitation and its partitioning into evapotranspiration, runoff and moisture, it can provide important constraints in upcoming model experiments. At the same time, efforts are made to further reduce uncertainties and discrepancies in the CMIP6 model ensemble by correcting output variables' biases, and developing spatial downscaling and data merging methods, which will probably improve also mTWS estimates. Next to traditional statistical methods[67,68] also deep learning methods are increasingly being applied[69]. Finally, we suggest that the continuation of GRACE/-FO will provide new opportunities for evaluating future CMIPs, along ideas outlined here. The MAGIC constellation[70,71], consisting of the NASA/DLR GRACE-C and ESA's NGGM satellite pair, would enable generating TWSA data with higher spatial and temporal resolution, shorter latency and improved error representation, and find its way into more data assimilation frameworks. At the same time, the community has begun to develop data sets to correct TWSA for effects (e.g., gravity changes due to large-reservoir operations) not included in ESM simulations, further facilitating the consistency between observations and models.

Methods

CMIP6 data processing

Coupled ESMs in the Climate Model Intercomparison Project phase 6 (CMIP6[33]) deliver information on climate changes from 1850 to 2100 for past and potential future developments of climate forces, e.g., greenhouse gas concentrations. In this study, we approximate CMIP6 simulated total water storage anomalies (mTWSA) by using the sum of the variables total soil moisture content (mrso) and surface snow amount (snw), the only two water storage-related variables provided by CMIP6 ESMs. The uncertainties of this approximation are hard to quantify, but qualitatively known. In contrast to observed TWSA, CMIP6-modeled TWSA neglect water storage changes occurring in surface waters and glaciers as well as anthropogenic groundwater abstractions and irrigation, while natural groundwater variability is only implicitly contained in total soil moisture content. Furthermore, structural model deficiencies regarding the representation of soil water dynamics, hydrological processes and land-atmosphere interactions contribute to uncertainties in CMIP6-derived TWSA. However, regions particularly affected by surface water or glacier mass changes and groundwater abstraction are, e.g., identified in the supplementary material of Jensen et al.[28] and only have a marginally overlap with the regions discussed in this study.

Concatenating corresponding monthly output of the historical experiments for 1950–2014 and the experiments for the future scenario SSP5-8.5 for 2015–2100 provides a continuous time series of 150 years of mTWSA. A uniform spatial resolution was achieved by re-mapping each of the model output time series to a 2° grid. Subsequently, the linear trend is calculated by fitting a function

$$f(t) = a + b \cdot t + c \cdot cos(\omega t) + d \cdot sin(\omega t) + e \cdot cos(2\omega t) + f \cdot sin(2\omega t) \tag{1}$$

with parameters for bias (a), linear trend (b), annual and semi-annual cycle (c,d,e,f) to the time series by means of least squares adjustment. When the time series covers full years (as in our study), the estimated trend is equal to fitting a pure linear function (except for the former estimation being numerically more stable due to usually smaller residuals). If the time series does not cover full years, the estimated trend may be distorted if the annual/semi-annual cycle is not co-estimated. The standard deviation of the linear trend (b) is estimated from the postfit residuals. Based on this standard deviation the significance ($\alpha = 5\%$) of the trend is computed.

In total, 25 individual model experiments are providing the variables mrso and snw in the CMIP6 archive by the time of writing. Details about the individual models (including references and information on sub-models) can be found, e.g., in Table AII.5 in Annex II of the Sixth Assessment Report of the Intergovernmental Panel on Climate Change[72]. The 25 models are not all fully independent from each other. Some of them are improvements or extensions of each other, or they share central elements, such as land, atmosphere, or ocean sub-models. In order to obtain unbiased results when analyzing multi-model averages, we reduce the ensemble by omitting all highly correlated experiments by calculating pair-wise correlations of the vectorized mTWSA trend maps, similarly to Jensen et al.[41] using a threshold of 0.7. The assessment in this study is based on the remaining 17 individual coupled-climate models with altogether 105 ensemble members.

The multi-model median (MMMed) trend is obtained by calculating the median of all $N = 105$ individual mTWS trend values for each grid cell excluding Greenland, Svalbard, and Antarctica from the analysis. In order to give each model the same weight, regardless of the number of ensemble members belonging to it, we compute the weighted multi-model median as described in Jensen et al.[28]. However, the median operator smooths out extreme values in each grid cell, resulting in the range of values in the MMMed trend grid becoming smaller than the actual range of values in individual ensemble members. We therefore use a rescaled MMMed trend with rescaling factors based on the empirical cumulative density functions (ECDFs) of the MMMed and compared to the mean ECDF of the individual ensemble members. Details of this procedure can be found in Jensen et al.[28].

Computation of model consensus and temporal consensus maps

To identify regions in which the ensemble of CMIP6 ESMs has a high agreement about wetting/drying trends, we compute the so-called model consensus as the number of models that agree on the sign of the trend for each grid cell. Further analysis is then restricted to regions with high model consensus (i.e., at least 75% of the models agreeing on the sign of the trend).

To investigate the agreement of short-term trends with the long-term centennial trends, we perform a model study: We cut the 150-year time series into (overlapping) slices of 17-year time series (starting every 5 years) and calculate a 17-year trend map for each slice. Afterward, from all slices, we compute for each grid cell the percentage of 17-year trends that match the 150-year trend. This procedure is repeated for 20, 25, 30, ..., 50 years. The corresponding maps give an indication in which regions the agreement with the long-term trend is comparably high for each individual time span length. To synthesize this information, we additionally compute a map of high temporal consensus marking the grid cells that have an above-average trend agreement with the centennial trend for all differently long time span lengths.

Observational data

As observation-based data sets we use two TWSA data sets. The first one is derived from the satellite missions GRACE and GRACE-FO and the second one is the GLWS2.0 data set[43], referred to as GRACE/DA here, a synthesized data set derived by assimilating GRACE/-FO TWSA into a global hydrological model.

The GRACE satellite mission is a gravity mission by NASA/DLR that was launched in 2002 and ended in 2017[10]. The mission consisted of two satellites flying in the same orbit and used a K-band range measurement constellation between the two satellites to derive global gravity field measurements. The successor GRACE-FO was launched in 2018 and is additionally equipped with a laser-ranging interferometer to derive a higher accuracy of the inter-satellite distance[73].

To derive TWSA time series from the GRACE and GRACE-FO data, the monthly gravity fields are processed by the IGG standard processing: ITSG2018 operational Level 2 gravity field coefficients are provided by TU Graz and used in the further processing. Lower degree coefficients are replaced by data from other sources because GRACE/-FO measure them inprecisely[74,75] (received from http://grace.jpl.nasa.gov) and a DDK3 filter is applied to account for striping errors[76,77]. Then, a temporal mean is reduced and the Level 2 coefficients are transformed via spherical harmonic synthesis to the required gridded Level 3 total water storage anomalies. Typically, the final grid resolution is 0.5°, here we transform the TWSA to 2° grids to enable direct comparison with CMIP6 model data. The effect of glacial isostatic adjustment (GIA) is removed by using a model[78].

In case of comparing original GRACE/-FO data to other data sets or model simulations, the processing is finished at this stage. For the assimilation of GRACE-/FO into a model, we need to additionally aggregate the standard 0.5° TWSC grids to 4° grids to account for numerical issues[79]. It is important to understand that the 4° spatial resolution serves only as input to the data assimilation framework. The output of the data assimilation and thus the resolution of GRACE/DA is similar to the native model resolution, which is 0.5°.

GRACE/DA is a global product that provides total water storage anomalies, groundwater, soil moisture, snow, and surface water with a spatial resolution of 0.5° from 2003 to 2019. The data set derives these outputs by assimilating GRACE/-FO total water storage anomalies into the WaterGAP hydrology model (WGHM[45]). The assimilation framework is based on previous work of Eicker et al.[44] and Schumacher et al.[79], developed at the Institute of Geodesy and Geoinformation at the University of Bonn, Germany, and is updated by using the parallel data assimilation framework (PDAF[80]) and extending the regional to a global scale. It uses the Ensemble Kalman Filter[81] to sequentially update the predicted model states by integrating GRACE/-FO observations based on uncertainty information via ensembles. Thus, the resulting data set represents an optimal synthesis of the GRACE/-FO data and the data sets that went into the hydrological model. By assimilating GRACE/-FO into WGHM we spatially downscale the GRACE/-FO fields to the model's resolution and vertically disaggregate the TWSA into the different water compartments that WGHM represents. Since the model is based on uncertainty given in forcing data and model assumptions, the assimilation certainly improves the model's realism by pulling the simulations closer to the GRACE/-FO observations. The GRACE/DA data set was e.g., evaluated against another 0.5° global land water storage data set (CLSM-DA) derived from assimilating GRACE/-FO into the CLSM model, showing a good agreement of the linear trends in TWS[43]. Linear trends from the assimilation product further lie in-between the trends from the hydrological model and the GRACE/-FO observations used in the assimilation procedure. Furthermore, the GRACE/DA product was compared to in situ GNSS observations of vertical loading at 1000+ stations globally with the result that it correlates better at short-term, seasonal, and long-term (>1.4 yr) temporal bands than GRACE/-FO. We note that in the GNSS analysis, the linear trends had to be removed in advance (due to tectonic effects) to enable a comparison.

WGHM represents a global hydrological model that simulates water fluxes and storages on an 0.5° grid based on horizontal and vertical water routing; here we use version 2.2e. The model represents soil moisture via a single layer that extends to the root zone, it simulates varying surface water storages (lakes, wetlands, rivers, and reservoirs), and it includes a conceptual groundwater representation. Anthropogenic water use is included in WGHM with a specific focus on human water use (surface and groundwater abstractions) and man-made reservoirs[45]. The model is mainly forced by precipitation, temperature, longwave and shortwave radiation. The forcing data used here is derived by the GSWP3-W5E5 data set[82].

Data availability

The CMIP6 data are publicly available, e.g., via https://esgf-data.dkrz.de/search/cmip6-dkrz/. The GRACE/DA data used in this contribution can be downloaded from PANGAEA. Derived data that supports the findings of this study are available from the corresponding author L.J. on request.

Code availability

The Python codes for data analysis and figures generation are available upon request from the corresponding author L.J.

Received: 15 February 2024; Accepted: 25 September 2024;
Published online: 17 October 2024

References

1. Held, I. & Soden, B. Robust responses of the hydrological cycle to global warming. *J. Clim.* **19**, 5686–5699 (2006).
2. Chou, C., Neelin, J., Chen, C.-A. & Tu, J.-Y. Evaluating the rich-get-richer mechanism in tropical precipitation change under global warming. *J. Clim.* **22**, 1982–2005 (2009).
3. Greve, P. et al. Global assessment of trends in wetting and drying over land. *Nat. Geosci.* **7**, 716–721 (2014).
4. Durack, P., Wijffels, S. & Matear, R. Ocean salinities reveal strong global water cycle intensification during 1950 to 2000. *Science* **336**, 455–458 (2012).
5. Roderick, M. L., Sun, F., Lim, W. H. & Farquhar, G. D. A general framework for understanding the response of the water cycle to global warming over land and ocean. *Hydrol. Earth Syst. Sci.* **18**, 1575–1589 (2014).
6. Eicker, A., Forootan, E., Springer, A., Longuevergne, L. & Kusche, J. Does GRACE see the terrestrial water cycle "intensifying"? *J. Geophys. Res.* **21**, 733–745 (2016).
7. Xiong, J., Guo, S., Abishek, J. & Yin, J. Global evaluation of the "dry gets drier, and wet gets wetter" paradigm from a terrestrial water storage change perspective. *HESS* **26**, 6457–6476 (2022).
8. Greve, P. & Seneviratne, S. Assessment of future changes in water availability and aridity. *Geophys. Res. Lett.* **42**, 5493–5499 (2015).
9. Blöschl, G. et al. Changing climate both increases and decreases European river floods. *Nature* **573**, 108–111 (2019).
10. Tapley, B. D., Bettadpur, S., Watkins, M. & Reigber, C. The gravity recovery and climate experiment: mission overview and early results. *Geophys. Res. Lett.* **31**, L09607 (2004).
11. McGirr, R., Tregoning, P., Purcell, A. & McQueen, H. Significant local sea level variations caused by continental hydrology signals. *Geophys. Res. Lett.* **51**, e2024GL108394 (2024).
12. Famiglietti, J. & Rodell, M. Water in the balance. *Science* **340**, 1300–1301 (2013).
13. Cazenave, A., WCRP Global Sea Level Budget Group. Global sea-level budget 1993-present. *ESSD* **10**, 1551–1590 (2018).
14. Frederikse, T. et al. The causes of sea-level rise since 1900. *Nature* **584**, 393–397 (2020).
15. Rodell, M., Velicogna, I. & Famiglietti, J. S. Satellite-based estimates of groundwater depletion in India. *Nature* **460**, 999–1002 (2009).
16. Voss, K. A. et al. 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 (2013).
17. Liu, M., Pei, H. & Shen, Y. Evaluating dynamics of GRACE groundwater and its drought potential in Taihang Mountain Region, China. *J. Hydrol.* **612**, 128156 (2022).

18. Abdelmohsen, K. et al. Buffering the impacts of extreme climate variability in the highly engineered Tigris Euphrates river system. *Sci. Rep.* **12**, 4178 (2022).
19. Scanlon, B. R., Jolly, I., Sophocleous, M. & Zhang, L. Global impacts of conversions from natural to agricultural ecosystems on water resources: Quantity versus quality. *Water Resour. Res.* **43**, W03437 (2007).
20. Chang, L.-L. & Niu, G.-Y. The impacts of interannual climate variability on the declining trend in terrestrial water storage over the Tigris-Euphrates river basin. *J. Hydrometeorol.* **24**, 549–560 (2023).
21. Rodell, M. et al. Emerging trends in global freshwater availability. *Nature* **557**, 651 (2018).
22. Scanlon, B. R. et al. Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. *Proc. Natl Acad. Sci. USA* **115**, E1080–E1089 (2018).
23. Kornfeld, R. P. et al. GRACE-FO: the gravity recovery and climate experiment follow-on mission. *J. Spacecr. Rockets* **56**, 931–951 (2019).
24. Tapley, B. D. et al. Contributions of GRACE to understanding climate change. *Nat. Clim. Change* **9**, 358–369 (2019).
25. Landerer, F. W. et al. Extending the global mass change data record: GRACE Follow-On instrument and science data performance. *Geophys. Res. Lett.* **47**, e2020GL088306 (2020).
26. Springer, A., Eicker, A., Bettge, A., Kusche, J. & Hense, A. Evaluation of the water cycle in the European COSMO-REA6 reanalysis using GRACE. *Water* **9**, 289 (2018).
27. Wu, R.-J., Lo, M.-H. & Scanlon, B. The annual cycle of terrestrial water storage anomalies in CMIP6 models evaluated against GRACE data. *J. Clim.* **34**, 8205–8217 (2021).
28. Jensen, L., Eicker, A., Dobslaw, H. & Pail, R. Emerging changes in terrestrial water storage variability as a target for future satellite gravity missions. *Remote Sens.* **12**, 3898 (2020a).
29. Zhao, M., A, G., Velicogna, I. & Kimball, J. A global gridded dataset of GRACE drought severity index for 2002-14: Comparison with PDSI and SPEI and a case of the Australia millenium drought. *J. Hydrometeorol.* **18**, 2117–2129 (2017).
30. Gerdener, H., Engels, O. & Kusche, J. A framework for deriving drought indicators from the gravity recovery and climate experiment (GRACE). *HESS* **24**, 227–248 (2020).
31. Haagmans, R., Siemes, C., Massotti, L., Carraz, O. & Silvestrin, P. ESA's Next Generation Gravity Mission concepts. *Rend. Lincei. Sci. Fis. Nat.* **31**, S15–S25 (2020).
32. Deggim, S. et al. RECOG RL01: correcting GRACE total water storage estimates for global lakes/reservoirs and earthquakes. *Earth Syst. Sci. Data* **13**, 2227–2244 (2021).
33. Eyring, V. et al. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.* **9**, 1937–1958 (2016).
34. Fiedler, S. et al. Simulated tropical precipitation assessed across three major phases of the Coupled Model Intercomparison Project (CMIP). *Monthly Weather Rev.* **148**, 3653–3680 (2020).
35. Bock, L. et al. Quantifying progress across different CMIP phases with the ESMValTool. *J. Geophys. Res.: Atmospheres* **125**, e2019JD032321 (2020).
36. Wang, D. & Zeng, Z. Urgent need to improve modelled sensitivity of evaporation to vegetation change. *Nat. Water* **2**, 211–214 (2024).
37. Qiao, Y. et al. Contrasting sensitivity of air temperature trends to surface soil temperature trends between climate models and reanalyses. *NPJ Clim. Atmos. Sci.* **7**, 1–11 (2024).
38. Qiao, L., Zuo, Z. & Xiao, D. Evaluation of soil moisture in CMIP6 simulations. *J. Clim.* **35**, 779–800 (2022).
39. Ossó, A., Craig, P. & Allan, R. P. An assessment of CMIP6 climate signals and biases in temperature, precipitation and soil moisture over Europe. *Int. J. Climatol.* **43**, 5698–5719 (2023).
40. Hsu, H. & Dirmeyer, P. A. Uncertainty in projected critical soil moisture values in CMIP6 affects the interpretation of a more moisture-limited world. *Earth's. Future* **11**, e2023EF003511 (2023).
41. Jensen, L., Eicker, A., Dobslaw, H., Stacke, T. & Humphrey, V. Long-term wetting and drying trends in land water storage derived from GRACE and CMIP5 models. *J. Geophys. Res.: Atmospheres* **124**, 9808–9823 (2019).
42. Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93**, 485–498 (2011).
43. Gerdener, H., Kusche, J., Schulze, K., Döll, P. & Klos, A. The global land water storage data set release 2 (GLWS2. 0) derived via assimilating GRACE and GRACE-FO data into a global hydrological model. *J. Geod.* **97**, 73 (2023).
44. Eicker, A., Schumacher, M., Kusche, J., Döll, P. & Müller Schmied, H. 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 (2014).
45. Müller Schmied, H. et al. The global water resources and use model WaterGAP v2.2d: model description and evaluation. *Geosci. Model Dev.* **14**, 1037–1079 (2021).
46. Li, B. et al. Global GRACE data assimilation for groundwater and drought monitoring: advances and challenges. *Water Resour. Res.* **55**, 7564–7586 (2019).
47. Chen, J., Wilson, C., Tapley, B., Save, H. & Cretaux, J.-F. Long-term and seasonal Caspian Sea level change from satellite gravity and altimeter measurements. *J. Geophys. Res.: Solid Earth* **122**, 2274–2290 (2017).
48. Getirana, A. Extreme water deficit in Brazil detected from space. *J. Hydrometeorol.* **17**, 591–599 (2016).
49. Panda, D. K. & Wahr, J. Spatiotemporal evolution of water storage changes in India from the updated GRACE-derived gravity records. *Water Resour. Res.* **52**, 135–149 (2016).
50. Luthcke, S. B. et al. Antarctica, Greenland and Gulf of Alaska land-ice evolution from an iterated GRACE global mascon solution. *J. Glaciol.* **59**, 613–631 (2013).
51. Gardner, A. S. et al. Sharply increased mass loss from glaciers and ice caps in the Canadian Arctic Archipelago. *Nature* **473**, 357–360 (2011).
52. Swenson, S. & Lawrence, D. Assessing a dry surface layer-based soil resistance parameterization for the community land model using GRACE and FLUXNET-MTE data. *J. Geophys. Res. Atmosphres* **119**, 10,299–10,312 (2014).
53. Vereecken, H. et al. Soil hydrology in the Earth system. *Nat. Rev. Earth Environ.* **3**, 573–587 (2022).
54. Koster, R. D. et al. On the nature of soil moisture in land surface models. *J. Clim.* **22**, 4322–4335 (2009).
55. Cos, J. et al. The Mediterranean climate change hotspot in the CMIP5 and CMIP6 projections. *Earth Syst. Dyn.* **13**, 321–340 (2022).
56. Chen, Z. et al. Global land monsoon precipitation changes in CMIP6 projections. *Geophys. Res. Lett.* **47**, e2019GL086902 (2020).
57. Pelletier, J. D. et al. A gridded global data set of soil, intact regolith, and sedimentary deposit thicknesses for regional and global land surface modeling. *J. Adv. Model. Earth Syst.* **8**, 41–65 (2016).
58. Pokhrel, Y. N., Fan, Y. & Miguez-Macho, G. Potential hydrologic changes in the Amazon by the end of the 21st century and the groundwater buffer. *Environ. Res. Lett.* **9**, 084004 (2014).
59. Eicker, A. et al. Influence of GIA uncertainty on climate model evaluation with GRACE/GRACE-FO satellite gravimetry data. *J. Geophys. Res.: Solid Earth* **129**, e2023JB027769 (2024).
60. Zhong, X., Zhang, T., Kang, S. & Wang, J. Snow depth trends from CMIP6 models conflict with observational evidence. *J. Clim.* **35**, 1293–1307 (2022).
61. van Dijk, A. I. J. M. et al. The Millennium Drought in southeast Australia (2001-2009): natural and human causes and implications for water resources, ecosystems, economy, and society. *Water Resour. Res.* **49**, 1040–1057 (2013).

62. Crosbie, R., Wang, B., Kim, S., Mateo, C. & Vaze, J. Changes in the surface water - Groundwater interactions of the Murray-Darling basin (Australia) over the past half a century. *J. Hydrol.* **622**, 129683 (2023).
63. Zarfl, C., Lumsdon, A. E., Berlekamp, J., Tydecks, L. & Tockner, K. A global boom in hydropower dam construction. *Aquat. Sci.* **77**, 161–170 (2015).
64. Scanlon, B. R. et al. Linkages between GRACE water storage, hydrologic extremes, and climate teleconnections in major African aquifers. *Environ. Res. Lett.* **17**, 014046 (2022).
65. Löcher, A. & Kusche, J. A hybrid approach for recovering high-resolution temporal gravity fields from satellite laser ranging. *J. Geod.* **95**, 1–15 (2021).
66. Almazroui, M. et al. Assessment of CMIP6 performance and projected temperature and precipitation changes over South America. *Earth Syst. Environ.* **5**, 155–183 (2021).
67. Shrestha, R. R., Schnorbus, M. A., Werner, A. T. & Zwiers, F. W. Evaluating hydroclimatic change signals from statistically and dynamically downscaled GCMs and hydrologic models. *J. Hydrometeorol.* **15**, 844–860 (2014).
68. Miao, C., Su, L., Sun, Q. & Duan, Q. A nonstationary bias-correction technique to remove bias in GCM simulations. *J. Geophys. Res.: Atmospheres* **121**, 5718–5735 (2016).
69. Wei, X. et al. Deep-learning-based harmonization and super-resolution of near-surface air temperature from CMIP6 models (1850-2100). *Int. J. Climatol.* **43**, 1461–1479 (2023).
70. Massotti, L., Siemes, C., March, G., Haagmans, R. & Silvestrin, P. Next generation gravity mission elements of the mass change and geoscience international constellation: From orbit selection to instrument and mission design. *Remote Sens.* **13**, 3935 (2021).
71. Daras, I. et al. Mass-change And Geosciences International Constellation (MAGIC) expected impact on science and applications. *Geophys. J. Int.* **236**, 1288–1308 (2024).
72. Intergovernmental Panel on Climate Change (IPCC) (ed.). In *Climate Change 2021 - The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, 2087–2138, https://doi.org/10.1017/9781009157896.016 (Cambridge University Press, 2023).
73. Abich, K. et al. In-orbit performance of the GRACE follow-on laser ranging interferometer. *Phys. Rev. Lett.* **123**, 031101 (2019).
74. Swenson, S., Chambers, D. & Wahr, J. Estimating geocenter variations from a combination of GRACE and ocean model output. *J. Geophys. Res.: Solid Earth* **113**, B08410 (2008).
75. Cheng, M., Ries, J. C. & Tapley, B. D. Variations of the Earth's figure axis from satellite laser ranging and GRACE. *J. Geophys. Res.: Solid Earth* **116**, B01409 (2011).
76. Kusche, J. Approximate decorrelation and non-isotropic smoothing of time-variable GRACE-type gravity field models. *J. Geod.* **81**, 733–749 (2007).
77. Kusche, J., Schmidt, R., Petrovic, S. & Rietbroek, R. Decorrelated GRACE time-variable gravity solutions by GFZ, and their validation using a hydrological model. *J. Geod.* **83**, 903–913 (2009).
78. Geruo, A., Wahr, J. & Zhong, S. Computations of the viscoelastic response of a 3-D compressible Earth to surface loading: an application to Glacial Isostatic Adjustment in Antarctica and Canada. *Geophys. J. Int.* **192**, 557–572 (2012).
79. Schumacher, M., Kusche, J. & Döll, P. A systematic impact assessment of GRACE error correlation on data assimilation in hydrological models. *J. Geod.* **90**, 537–559 (2016).
80. Nerger, L. & Hiller, W. Software for ensemble-based data assimilation systems-implementation strategies and scalability. *Comput. Geosci.* **55**, 110–118 (2013).
81. Evensen, G. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dyn.* **53**, 343–367 (2003).
82. Lange, S., Mengel, M., Treu, S. & Büchner, M. ISIMIP3a atmospheric climate input data (v1. 0), ISIMIP repository [data set], https://doi.org/10.48364/ISIMIP.982724 (2022).

Acknowledgements

We acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP6. We thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and ESGF. We acknowledge N.J. Steinert for providing the information of total soil depth for several CMIP6 models. J.K. and S.F. acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for SFB 1502/1-2022 (Projektnummer: 450058266). H.G. acknowledges funding from DFG within the RU GlobalCDA (KU1207/26-1).

Author contributions

L.J. and A.E. conceived the experiments, L.J. conducted the experiments, H.G. created and provided the GRACE/DA data set, J.K. and S.F. analyzed the results in selected regions. All authors contributed to the text and reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Laura Jensen.

Reprints and permissions information is available at http://www.nature.com/reprints