JGR Solid Earth

RESEARCH ARTICLE
10.1029/2024JB031011

Return Levels of Dry Extreme Events in Terrestrial Water Storage From Satellite Gravimetry and CMIP6 Global Coupled Climate Models

Klara Middendorf[1], **Henryk Dobslaw**[2], **Laura Jensen**[2], **and Annette Eicker**[1]

[1]HafenCity University Hamburg, Hamburg, Germany, [2]GFZ Helmholtz Centre for Geosciences, Potsdam, Germany

Key Points:

- One-in-ten years dry return levels in terrestrial water storage obtained with generalized extreme value distributions for GRACE and CMIP6
- Strong deviations found in individual model runs indicate value of GRACE observations to inform numerical models
- CMIP6 ensemble projects an increase in the severity of dry return levels for the 21st century

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

K. Middendorf,
klara.middendorf@hcu-hamburg.de

Citation:

Middendorf, K., Dobslaw, H., Jensen, L., & Eicker, A. (2025). Return levels of dry extreme events in terrestrial water storage from satellite gravimetry and CMIP6 global coupled climate models. *Journal of Geophysical Research: Solid Earth*, *130*, e2024JB031011. https://doi.org/10.1029/2024JB031011

Received 17 DEC 2024
Accepted 19 SEP 2025

Author Contributions:

Conceptualization: Klara Middendorf, Henryk Dobslaw, Laura Jensen, Annette Eicker
Formal analysis: Klara Middendorf
Funding acquisition: Annette Eicker
Investigation: Klara Middendorf
Methodology: Klara Middendorf, Henryk Dobslaw, Laura Jensen, Annette Eicker
Project administration: Annette Eicker
Software: Klara Middendorf
Supervision: Annette Eicker
Visualization: Klara Middendorf
Writing – original draft: Klara Middendorf, Henryk Dobslaw, Annette Eicker

Abstract Satellite gravimetry as realized with GRACE and GRACE-FO provides a novel opportunity to study extreme deviations from annually varying terrestrial water storage (TWS) in all continental areas of our planet. By utilizing the generalized extreme value (GEV) distribution, we estimate return levels for events that are expected to happen once every 10 (i.e., 1-in-10) years. With two GRACE-like reconstructions spanning over 40 and 114 years, respectively, we show that the currently available data record of 20 years is already sufficiently long to derive robust estimates of those return levels. When contrasting the GRACE/-FO results to model experiments from the CMIP6 archive extending until the year 2100 by concatenating historical runs and climate projections under the SSP5-8.5 socioeconomic pathway, we find that (a) the multi-model median from CMIP6 has the overall best agreement with the satellite data, thereby nicely confirming the validity of a central assumption of many climate-related studies that heavily rely on ensemble statistics. We also find that (b) CMIP6 model runs contain only modest deviations of 1-in-10 years return levels from the beginning of the 20th century when compared to present-day, but predict stronger changes toward more extreme return levels by the end of the 21st century. On the other hand, we also find substantial differences between satellite data and individual model experiments, which opens new opportunities to inform, validate and/or calibrate numerical climate models with satellite gravimetry data from GRACE, GRACE-FO, and in future also GRACE-C.

Plain Language Summary The satellite mission GRACE and its successor, GRACE-FO, represent a unique observation tool that is able to globally capture changes in the distribution of water masses on and below the Earth's surface. Over land, these measurements of "Terrestrial Water Storage" (consisting of water stored in soil, groundwater, snow, ice, surface waters and vegetation) provide a novel way to validate numerical climate models, which are important for forecasting the future climate evolution of our planet. In this study, we use GRACE/-FO data to investigate the ability of climate models to simulate the frequency and intensity of extraordinary dry water storage conditions (dry extremes). While we find good agreement with GRACE/-FO in many regions when combining the results of 17 different climate models, there are stronger discrepancies between GRACE/-FO and individual climate models, which calls for further improvements of those particular models. We also find that over the 20th century the intensity of dry extremes has been rather stable, while it will become more extreme (i.e., drier) in the coming decades as long as global greenhouse gas emissions are not reduced drastically below the levels assumed in this study.

1. Introduction

Interactively coupled numerical climate models have been key tools to establish a solid understanding of various side-effects of rising greenhouse gas concentrations on global and regional climate change. It is widely accepted that the global water cycle is expected to intensify with increasing atmospheric temperature, a consequence well in line with the Clausius-Clapeyron relationship. The general conclusion "wet gets wetter, dry gets drier" has been put forward by many model-based studies (Chou et al., 2009; Held & Soden, 2006). Questions still remain, however, on the details of those changes—both in terms of increasing drought conditions and flood risks—for certain regions (Chen et al., 2020; Moon & Ha, 2020; X. Huang et al., 2024). In addition, results from individual numerical climate models differ substantially when reliable predictions for a particular planning horizon (e.g., 10, 30, or even 50 years) into the future are requested. It is therefore important to continue improving such models by both refining the representation of existing model dynamics and by adding new processes and feedbacks that were previously considered of secondary importance.

Writing – review & editing: Henryk Dobslaw, Laura Jensen, Annette Eicker

In addition to the modeling work, precise observations on the evolution of the water cycle are needed to assess the predictive capabilities of numerical models. A rich source of observational evidence arises from geodetic methods that are specifically well suited to monitor slow and subtle changes in the Earth system. A prominent example are the Gravity Recovery And Climate Experiment (GRACE) satellites (Tapley et al., 2019), which have been providing monthly mean snapshots of the water mass distribution on the continents since the launch of the first mission in the year 2002. After the end of the GRACE operation in 2017, the time series is currently being continued with GRACE-FO (Landerer et al., 2020), and preparations are well underway to launch with GRACE-C the third mission of this type presumably in the year 2028 (Flechtner et al., 2024). With further activities by other space agencies to initiate even more gravity satellites past the year 2030, it is safe to assume that such data will be also available in the medium-term future.

GRACE/-FO satellite observations have been used to quantify water storage change for a wide range of hydrometeorological applications. Döll et al. (2014) compared simulated groundwater depletion from the WaterGAP model with GRACE-derived estimates. Rodell et al. (2015) utilized GRACE data to quantify the current state of the global water and energy cycles. Experiments from different global hydrological models and land surface schemes under controlled atmospheric forcing have been evaluated against GRACE data in various ways (Scanlon et al., 2018; Zhang et al., 2016). Similar analyses have been also performed for coupled climate models by Jensen et al. (2019, 2020). GRACE data is also being used to assess the potential of future flooding by taking into account the pre-existing water storage prior to a major precipitation event (Reager & Famiglietti, 2009; Reager et al., 2014). Other studies have rather focussed on water deficits tracked by the satellites, which helped to quantify the extended use of water resources in heavily irrigated areas like California's Central Valley (Carlson et al., 2024; Scanlon et al., 2012) or the Ganges catchment in Northern India (Panda & Wahr, 2016; Tiwari et al., 2009; Vissa et al., 2019). Gerdener et al. (2020) compared water storage-related drought indicators and applied them for analyzing GRACE time series in, for example, South Africa. In such areas of high water demand, the availability of a satellite-based monitoring system is increasingly acknowledged also by governmental agencies that are tasked to manage existing water resources in a sustainable way by accommodating demands of the different interest groups involved.

GRACE data has been also previously utilized to access return levels of hydrometeorological extremes (Kusche et al., 2016). For rather short data records, however, it is theoretically challenging to derive reliable return levels, in particular if time intervals longer than the actual observation period are targeted. Statistical methods for the characterization of extremes were primarily tailored in the past towards temperature and precipitation extremes, two quantities for which long and extended observation records are available. This led to, for example, the 10 indicators described by Frich et al. (2002) or the Standardized Precipitation Index. Other studies have applied extreme value theory (EVT) to hydrometeorological data comprising of, e.g., stream flow (e.g., Kochanek et al., 2014; Morrison & Smith, 2002), surface temperature (e.g., García-Cueto et al., 2014; Xu, 2019), or precipitation (e.g., Chikobvu & Chifurira, 2015; Papalexiou & Koutsoyiannis, 2013). It was found that probability distributions of hydrological extremes are often heavy-tailed (Katz et al., 2002), whereas distributions of temperature extremes tend to be short-tailed (Nerantzaki & Papalexiou, 2022; J. Wang et al., 2016). Those results underline that the common assumption of a normal (i.e., Gaussian) distribution (which is the underlying assumption of many classical uncertainty analyses) is not appropriate here.

In this study, we assess 1-in-10 years return levels of TWS annual minima, which can be considered as indicative of drought-severity reappearing once in every decade by using EVT. We utilize the Coupled Model Intercomparison Project Phase 6 (CMIP6, Eyring et al., 2016) archive, which provides a range of coordinated numerical experiments from essentially all state-of-the-art numerical climate models from numerous international research groups. The models are contrasted against a combined GRACE/GRACE-FO data record covering 20 years, as well as two GRACE-like reconstructions that offer an extended data range of 40 and 114 years, respectively.

The paper is structured as follows: After introducing both models and data (Section 2), we discuss the extreme value theory applied in this study to identify TWS minima and the associated fitting of the probability density functions necessary to predict arbitrary return levels as the 1-in-10 years indicator selected here (Section 3). We then show return levels of TWS annual minima from GRACE/-FO and CMIP6 model-derived indicators, and discuss the possible impact of the rather short observational record by additionally utilizing much longer GRACE-like reconstructions (Section 4). Subsequently, we assess various options to characterize uncertainties for both the satellite observations and the numerical models for their respective fidelity to characterize extremes

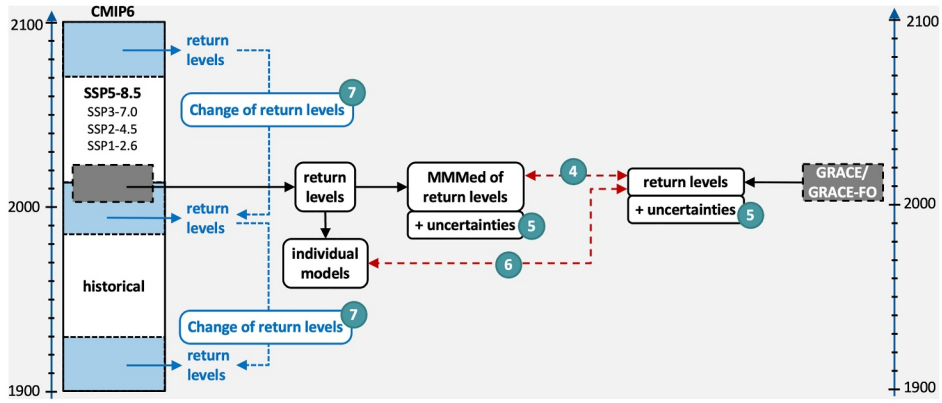

Figure 1. Overview of the methodology of this study. The green circles denote the chapters in which each analysis is addressed.

(Section 5). In Section 6, we identify CMIP6 models particularly well suited to predict dry extremes by evaluating individual climate model experiments. We also assess anticipated changes in dry extremes until the end of the current century as predicted by CMIP6 under the assumption of the Shared Socioeconomic Pathway SSP5-8.5 scenario (Section 7) and compare them to the respective changes under more optimistic emission scenarios. The article closes with a brief summary and some conclusions for future research (Section 8).

We aim to demonstrate with this work that satellite gravity from GRACE/-FO is a valuable—but currently underutilized—information source on hydrometeorological variability that is, available over a sufficiently long period already. We focus on hydrometeorological extreme events characterized by extended periods of below-average rainfall and assess how such events are represented in different climate models. We are also aiming to quantify uncertainties with the estimated return levels from both observations and numerical models, and finally assess future changes in the return levels under future climate scenarios as represented by four different socio-economic pathways considered in CMIP6. The workflow of our study is graphically summarized in Figure 1.

2. Data

2.1. GRACE and GRACE-FO Satellite Observations

We use global monthly mean gravity field models from ITSG-Grace2018 (Kvas et al., 2019; Mayer-Gürr et al., 2018) for the time period April 2002 to December 2022 as spherical harmonic coefficients up to degree and order 120. Degree-1 terms as provided by Sun et al. (2016) based on Swenson et al. (2008) are added to consider geocenter motion. The c_{20} coefficient is replaced from an appropriate Satellite Laser Ranging time series (Cheng & Ries, 2017). To exclude the effects of glacial isostatic adjustment, the ICE6G_D model (Peltier et al., 2018) is subtracted, and to account for the striping pattern resulting from spatially correlated noise, a DDK3 filter (Kusche, 2007) is applied. We note that other GRACE filtering options are available (e.g., Horvath et al., 2018; Swenson & Wahr, 2006; Wahr et al., 1998), but when aiming for similar filter characteristics in terms of the level of noise suppression and of signal preservation, then the choice of the specific type of filter only has a minor influence on our results. The resulting spherical harmonic coefficients are converted to equivalent water heights on a 2° × 2° longitude-latitude grid by

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

with λ and θ being the spherical coordinates, M and R denoting mass and radius of the Earth, $\rho_w = 1000\ \mathrm{kg/m^3}$ is the density of water, k'_n are the Load Love Numbers (Lambeck, 1988), c_{nm} denote the filtered spherical harmonic coefficients of gravitational potential and $Y_{nm}(\lambda,\phi)$ are the surface spherical harmonic functions.

Short gaps in the GRACE record arise from brief periods of short repeat orbit conditions that degraded the spatial coverage, and the occasional switch-off of instruments due to energy limitations caused by degrading batteries

toward the end of the GRACE mission life-time. Short gaps of individual months are linearly interpolated for our analysis. This does not lead to any adverse consequences on the return level estimations, since the estimation of the return levels uses annual minimum values only. The months June 2017 to June 2018 are excluded completely from this study as neither GRACE nor GRACE-FO were operational at this time.

2.2. GRACE-Like Reconstructions

The GRACE/-FO timespan of 20 years is a rather short time period to investigate extreme events, which occur very rarely by definition. Furthermore, statistical methods become more reliable as more data is available, so we expand our analysis to GRACE-like reconstructions of TWS as published by Humphrey and Gudmundsson (2019a). The GRACE-REC product consists of six data sets based on combinations of two GRACE releases (JPL and GSFC mascons) with three meteorological data sets (MSWEP, ERA5, and GWSP3). Each data set includes 100 ensemble members that characterize internal uncertainty. In our study we utilize monthly ensemble-means of two reconstructions (JPL-ERA5 and JPL-GSWP3). While the ERA5-based products (1979–2018) are often superior on a grid cell scale, GSWP3 is offering centennial-long reconstructions (i.e., 1901–2014). The choice against which of the two GRACE products the reconstruction is calibrated, seems to be of minor importance for model performance, and we just choose the JPL solution for its wider use in the community. It is further noted that the GRACE-REC time series do not include a seasonal cycle, but trends caused by decadal variability and long-term changes in precipitation are well preserved. Similar to the satellite observations, all data from GRACE-REC are remapped to a $2° \times 2°$ longitude-latitude grid.

2.3. CMIP6 Model Experiments

The Coupled Model Intercomparison Project (CMIP) of the World Climate Research Program develops a common frame for numerical climate simulation experiments. Its aim is to foster the comparability of modeling approaches of the various international research teams that focus on climate projections and decadal climate predictions. CMIP distinguishes between mandatory experiments (e.g., historical simulations) and optional experiments structured as individual Model Intercomparison Projects (MIPs) that are more specific to certain research questions (Eyring et al., 2016).

In the ScenarioMIP (O'Neill et al., 2016), which is part of the most recently completed sixth phase of CMIP (i.e., CMIP6), simulations are based on different future development scenarios of climate and societies. Those future evolutions are described in the five Shared Socioeconomic Pathways (SSPs) containing narratives for different socio-economic developments having implications for, e.g., greenhouse gas emissions and land use changes. The SSPs are combined with the expected level of radiative forcing in the year 2100 to derive a SSP forcing scenario as input for CMIP6 climate projections. The SSP5-8.5 scenario is the scenario with the most severe—and maybe the most realistic (Carvalho et al., 2022)—prognosis for the future climate evolution, with a pathway based on an energy intensive, fossil-based economy while assuming optimistic trends for education, health, economy, and the effectiveness of global institutions. In this study, we mainly use the SSP5-8.5 experiments (2015–2100) for the 21st century and combine them with historical simulations (1850–2015) to obtain long-term time series describing the evolution of water mass on land over a period of 250 years. The less severe emission pathways (SSP1-2.6, SSP2-4.5, SSP3-7.0) are additionally investigated for comparison in Section 7.2. In the following analysis, we only focus on the simulations for the 20th and 21st century, that is, 1900–2100. Experiments carried out by a single numerical model code (so-called model runs) can differ along four principal dimensions: realizations based on small random perturbations of the initial conditions (r); differences in the initialization method (i); perturbations of the physics considered by the model (p); and changes in the external forcing of the model run (f). Model runs can be identified by numbering these dimensions, for example, r1i1p1f1, and together they form the ensemble of an individual model, whereas the sum of all experiments from different models are termed the multi-model ensemble throughout this paper.

To approximate modeled terrestrial water storage (mTWS), we add surface snow amount (variable id: `snw`) and total soil moisture content (variable id: `mrso`), which are the only water storage-related variables routinely output by CMIP6 models. Please note that the amount of soil layers and the total soil depths over which the mass content of water is integrated vertically to derive the total soil moisture content, varies between different models (cf. e.g., for CMIP5, Berg et al., 2017). Additionally, other potential sources of changes in water mass associated with groundwater or surface water dynamics are not included in mTWS. At the time of writing, 37 models provide

Table 1
List of All Climate Models Used in This Study That Are Both Considered as Sufficiently Independent and do Provide `mrso` *and* `snw` *Outputs for the Entire Period 1850–2100 as Composed of Both Historical Runs and the Subsequent SSP5-8.5 Projections*

Model (number simulations)	Institution	References
ACCESS-ESM1-5 (40)	CSIRO	Ziehn et al. (2019a, 2019b)
BCC-CSM2-MR (1)	BCC	Xin et al. (2018, 2019)
CanESM5 (50)	CCCma	Swart et al. (2019a, 2019b)
CESM2 (3)	NCAR	Danabasoglu (2019a, 2019b)
CIESM (1)	THU	W. Huang (2019a, 2019b)
CNRM-CM6-1 (6)	CNRM-CERFACS	Voldoire (2018, 2019)
E3SM-1-0 (1)	E3SM-Project, UCSB	Bader et al. (2019, 2022) and Stevenson et al. (2023a, 2023b)
EC-Earth3 (8)	EC-Earth-Consortium	EC-Earth Consortium (2019a, 2019b)
GFDL-CM4 (1)	NOAA-GFDL	Guo et al. (2018a, 2018b)
GISS-E2-1-G (14)	NASA-GISS	NASA Goddard Institute for Space Studies (NASA/GISS) (2018, 2020)
IPSL-CM6A-LR (7)	IPSL	Boucher et al. (2018, 2019)
MIROC6 (50)	MIROC	Tatebe and Watanabe (2018) and Shiogama et al. (2019)
MPI-ESM1-2-LR (50)	MPI-M	Wieners et al. (2019a, 2019b)
MRI-ESM2-0 (6)	MRI	Yukimoto et al. (2019a, 2019b)
NorESM2-LM (1)	NCC	Seland et al. (2019a, 2019b)
TaiESM1 (1)	AS-RCEC	Lee and Liang (2019, 2020)
UKESM1-0-LL (5)	MOHC, NIMS-KMA	Tang et al. (2019), Good et al. (2019), and Shim et al. (2020a, 2020b)

monthly data for the variables `snw` and `mrso` that are available from both historical and SSP5-8.5 experiments. Because some of these models share common model components they cannot be perceived as independent. Especially different models developed by one and the same institution usually produce highly correlated results (Jensen et al., 2019, 2020). Therefore, we limit our analysis to one model per institution. In case several models are available, we choose the one complying best with the criteria "highest number of ensemble members," "most basic degree of specialization," and "spatial resolution closest to $2° \times 2°$."

Thus, we utilize 17 different models with a total amount of 245 ensemble members for the time period of 1900–2100 (see Table 1). All grids are remapped to a $2° \times 2°$ spatial sampling. Since ice mass changes on land are not covered by most climate models, we exclude the areas of Greenland, Svalbard and Antarctica in our analysis.

We note that there are some differences in the quantities of TWS and mTWS (Jensen et al., 2019, 2020). As mentioned, mass changes caused by changes in glaciers, surface waters and groundwater as well as anthropogenic water use are not explicitly modeled in CMIP6. However, groundwater changes can be implicitly contained in total soil moisture content. GRACE/-FO satellites also observe non water-related mass changes on the continents —like residual GIA effects (Eicker et al., 2024), pre- and postseismic signals caused by large earthquakes (Chao & Liau, 2019; Panet et al., 2022) or residual atmospheric mass variability (Hardy et al., 2017)—that might superimpose the TWS signal. However, linear (e.g., GIA) or seasonal impacts do not affect our analysis. Additionally, many of these processes (e.g., glaciers, surface waters) affect the GRACE/-FO derived TWS signal only in isolated regions (see supplementary material in Jensen et al. (2020)).

It must be noted that despite the huge efforts made by the CMIP6 team to harmonize the database, not all experiment results can be utilized directly. For example, the simulation r1i1p1f1 of the CIESM model (downloaded from the CMIP6 archive on 18th July 2024) has `snw` and `mrso` data that is, two magnitudes smaller than other climate models. Since a similar issue was reported for precipitation some years ago, we decided to apply the factor 100 to mTWS derived by this model.

3. Extreme Values in Time Series and Derived Measures of Intensity

3.1. Extreme Value Theory

Extreme value theory (EVT) (cf. e.g., Coles, 2001; Leadbetter et al., 1983) is based on the distribution of the maxima M_n of a sequence of independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ following the distribution function $F(x)$

$$M_n = \max(X_1, \ldots, X_n). \tag{2}$$

Thus, the distribution function of M_n results as

$$\begin{aligned} G(x) = P(M_n \le x) &= P(X_1 \le x, \ldots, X_n \le x) \\ &= P(X_1 \le x) \cdot \ldots \cdot P(X_n \le x) \\ &= F^n(x). \end{aligned} \tag{3}$$

In many cases, the underlying parent distribution function $F(x)$ is unknown and can therefore not be used to derive $G(x)$. The Extremal Types Theorem (Fisher & Tippett, 1928; Gnedenko, 1943) says that if for some normalizing constants $a_n > 0, b_n$

$$P\left(\frac{M_n - b_n}{a_n} \le x\right) \to G(x) \tag{4}$$

holds, as $n \to \infty$, with $G(x)$ being a non-degenerate distribution, $G(x)$ must have the form of either the Gumbel, Fréchet, or Weibull distribution. These three distributions can be incorporated into the generalized extreme value (GEV) distribution via the following equation

$$G(x) = \begin{cases} \exp\left\{-\left(1 + \xi \cdot \frac{x-\mu}{\sigma}\right)^{-\frac{1}{\xi}}\right\} & \text{for } \xi \neq 0 \\ \exp\left\{-\exp\left\{\frac{x-\mu}{\sigma}\right\}\right\} & \text{for } \xi = 0. \end{cases} \tag{5}$$

The three parameters μ, σ, ξ denote the location, scale, and shape, respectively. The shape parameter runs the decay of the tail of the distribution and determines the distribution-type, with $\xi = 0$ corresponding to the Gumbel, $\xi > 0$ to the Fréchet, and $\xi < 0$ to the Weibull distribution. The GEV distribution represents the class of max-stable distribution functions. This means, that if some i.i.d. random variables follow a GEV distribution, the distribution of the maxima of those random variables is of the same type (i.e., has the same shape parameter) as the parent GEV distribution

$$G^n_{\xi,\mu,\sigma}(x) = G_{\xi,\mu',\sigma'}(x). \tag{6}$$

Please note that location and scale parameter are different, but convertible.

This limiting distribution of M_n can be exploited to derive an estimation of the distribution of the maxima, and thus the more extreme values, of an observed quantity. In many application cases associated with risk management, the goal is to deduce how often values in a certain order of magnitude occur (see Castillo et al., 2005, for examples of applications). This magnitude-frequency relationship of extreme values can be derived from the cumulative distribution function.

The return period R represents the time interval that elapses statistically until a value higher than a certain level occurs again

$$P(M_n > x_T) = \frac{1}{R}. \tag{7}$$

This level x_T is called the return level. With

$$G(x_T) = P(M_n \leq x_T) = 1 - \frac{1}{R} \tag{8}$$

it can be quantified as

$$\begin{aligned} x_T &= G^{-1}\left(1 - \frac{1}{R}\right) \\ &= \mu + \frac{\sigma}{\xi}\left(\left(-\ln\left(1 - \frac{1}{R}\right)\right)^{-\xi} - 1\right). \end{aligned} \tag{9}$$

The domain of R is $(1, \infty)$, so the temporal resolution is limited to one year, which means that for example, two extremes in one year cannot be detected (cf. C.-H. Wang & Holmes, 2020).

3.2. Block Maxima Method for TWS Return Levels

To apply the theoretical concept of random variables described above to a series of observations, the data is grouped into blocks of uniform length n. The maximum—or rather the minimum in our case—is then detected from each block, resulting in a series of extreme values for which the GEV distribution is assumed to be valid. After fitting the distribution into the extreme values, the return level for a certain return period can be calculated (e.g., Coles, 2001). In a hydrological context, it is common to choose a block length of 1 year to exclude any potential effects of the seasonal cycle.

It is important to emphasize that the assumption of i.i.d. data is often not justified in reality. Leadbetter et al. (1983) showed that the theory still applies for stationary sequences of random variables when assuming a common distribution F. Additionally, a distributional mixing condition $D(u_n)$ must hold. This condition states that the distribution function of maxima of separated sub-sequences are asymptotically independent with an increasing distance between them and implies a degree of independence between maxima of separated sub-sequences. Assuming that $G(x)$ denotes the limiting distribution of maxima of a sequence of i.i.d. random variables, as in Equation 4, the maxima M_n of a dependent stationary sequence of random variables that satisfies condition $D(u_n)$ and has the same distribution F can be expressed as

$$P(M_n \leq u_n) \to G^{\theta}(x) \tag{10}$$

with $\theta \in [0, 1]$ being called the extremal index. Since $G(x)$ is a max-stable distribution function, $G^{\theta}(x)$ is of the same type as $G(x)$ and therefore differs only in the scale and location parameter. Considering Equation 3, this connection can be interpreted as

$$F^{n\theta}((x + b_n)\, a_n) \to G^{\theta}(x) \tag{11}$$

when $n \to \infty$, which shows that the allowance of dependency reduces the number of (independent) random variables, of which a maximum is drawn, from n to θn. Thus, the effective sample size is reduced, which decreases the uncertainty of the GEV family as an approximation for the distribution of block maxima (Coles, 2001; Reiss & Thomas, 2007). When using monthly data, the usage of annual minima results in $n \leq 12$. Thus, the applicability of the asymptotic GEV distribution for $n \to \infty$ might be limited. However, even if the distribution for maxima of a finite sample deviates from its limiting distribution, it might be covered by one of the other types (i.e., other shape parameter) of the GEV distribution (Davison & Huser, 2015).

In order to come closer to the assumptions of stationarity and long-term temporal independence in TWS and mTWS data, a linear trend as well as annual and semi-annual harmonics were estimated and reduced from the data. As GRACE-REC does not include a seasonal cycle, only a linear trend was removed here. This procedure results in time series where minima do not necessarily represent droughts but rather extreme deviations from the expected conditions (i.e., this includes extraordinary dry wet periods as well). Hence, the investigated extremes in

this study characterize the variability of the data. We also note that minima in the years 2002, 2017, and 2018 were excluded in this study due to larger gaps in the GRACE/-FO data record.

3.3. Fitting the Distribution Function

There are different approaches for fitting the distribution function into a series of maxima. Most common are the method of moments (MOM); L-moments (LM) or the equivalent probability weighted moments (PWM); and maximum-likelihood estimators (MLE) that each come with different advantages and short-comings. Although MLE can be used to easily incorporate covariates (like the time) and allows to obtain error bounds more straightforwardly (Hamdi et al., 2018; Wilks, 2011), it can lead to numerical difficulties in the optimization process (Gubareva & Gartsman, 2010) and was shown to generate unrealistic values of the shape parameter for small samples and very short-tailed distributions (for $\xi < -0.5$) (Martins & Stedinger, 2000).

LM is computationally simple, shows a good performance for small samples and heavy-tailed distributions and is less influenced by outliers. Therefore, it is often used in hydrological applications (Früh et al., 2010; Katz et al., 2002; Wilks, 2011). It restricts the shape parameter to $\xi \leq 1$, whereas MOM limits it to $\xi \leq \frac{1}{3}$. Additionally MOM requires an iterative solution for ξ and often produces less accurate results (Früh et al., 2010; Martins & Stedinger, 2000). Due to its computational simplicity and better performance for small sample sizes, LM was chosen in this study. In addition, MOM was implemented for comparison and the derived return levels were found to be consistent with those derived using LM. For completeness, we note that there are also further approaches trying to overcome those short-comings, for example, a penalized maximum-likelihood estimator (Martins & Stedinger, 2000) or Bayesian methods (e.g., Stephenson & Tawn, 2004). A thorough review of those modern theoretic concepts is provided by Nerantzaki and Papalexiou (2022).

3.4. Uncertainty Assessment

The uncertainty of return levels can be split into two parts: (a) modeling uncertainty, referring to the uncertainty when fitting the parameters of a distribution function, and (b) statistical uncertainty, referring to the suitability of the asymptotic distribution function type itself (Li et al., 2014). The latter is difficult to assess, as the parent distribution of the data is unknown and therefore the true distribution of (independent) minima cannot be determined either. This uncertainty increases for return levels with higher return periods, especially when an extrapolation is performed (Li et al., 2014).

The modeling uncertainty depends on the used fitting technique as they result in different biases and uncertainties (cf. Section 3.3). An uncertainty assessment tool applicable for LM-fitted GEV parameters are bootstrapping techniques. Kyselý (2008) compares parametric and non-parametric bootstrap approaches and concludes that the parametric bootstrap outperforms the non-parametric bootstrap in most cases and should be preferred especially for small sample sizes and when a reasonable model for the data can be assumed. Still, Mudelsee (2010) points out, that this approach assumes a (somewhat) incorrect distribution model which should widen the confidence interval (CI) by an unknown extent. Kyselý (2008) also compares three different approaches for building a CI from the bootstrap samples and shows that the most simple CI, the percentile CI, is only inferior to the more advanced t-CIs and BCa-CIs when using large sample sizes, fitting of an incorrect distribution, or investigating small return periods. With those limitations in mind, we chose to perform an uncertainty assessment of the derived return levels by using parametric bootstrapping and building a percentile CI. However, this can only be regarded as a rough estimate of the uncertainty.

4. Return Levels of TWS Annual Minima From GRACE/-FO and CMIP6

In the following, we compare the return levels of TWS annual minima for a return period of 10 years. The longer the chosen return period, the bigger are the differences in the return level values, as slight deviations in the shape of the GEV distributions are most pronounced in the decay of their tails. On the other hand, the uncertainty of the distribution function itself increases strongly with a higher return period as, the scarcer the extremes of interest are, the less data is available to support the distribution curve in this value range. We therefore limit the discussion in this paper to the return levels of events that are statistically exceeded once every decade. For reasons of linguistic clarity, we will always refer to absolute values of TWS deficits throughout the text of this paper.

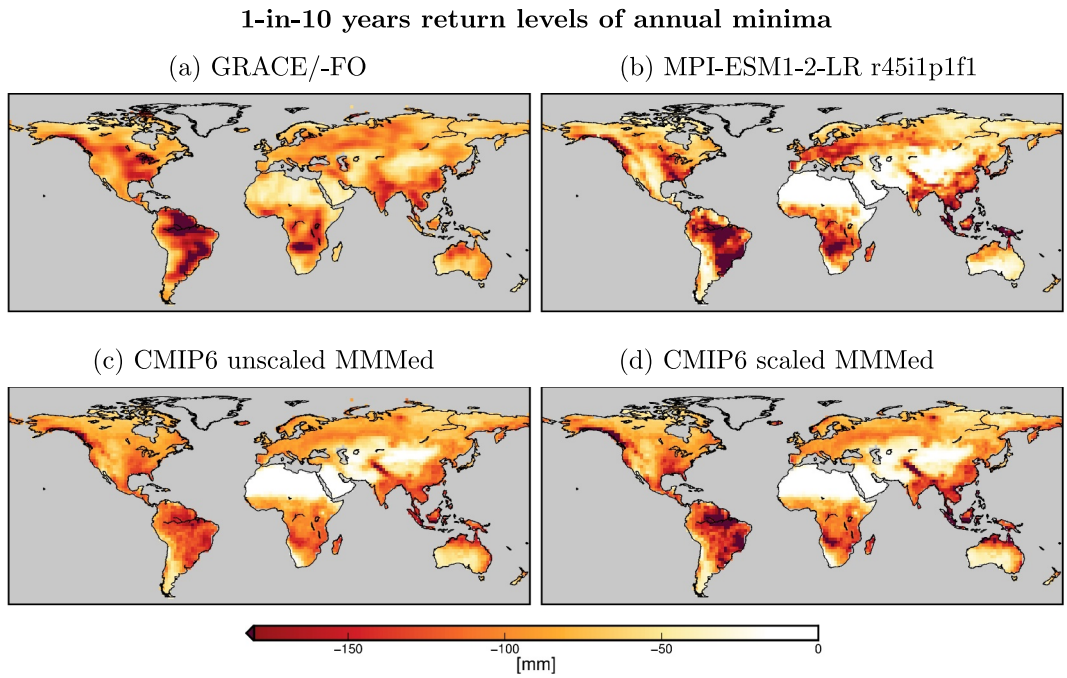

Figure 2. One-in-ten years return levels of TWS annual minima from 2002 to 2022 from (a) GRACE/GRACE-FO satellite observations; (b) a single realization (r45i1p1f1) of the MPI-ESM1-2-LR contribution to CMIP6; (c) the unscaled multi-model median from CMIP6; and (d) the scaled multi-model median from CMIP6.

Figure 2a shows the 1-in-10 years return levels of TWS annual minima derived from GRACE/-FO globally. We note deficits exceeding 150 mm of equivalent water height (ewh) in the Amazon, Orinoco and Parana catchments in South America, and in the upper tributaries of the Zambezi river in South Africa. We also note moderately high return levels of more than 100 mm in the Mid-Western States of the U.S. and central China. Return levels of roughly 30–60 mm in arid areas like the Sahara, however, rather reflect the current noise level of the GRACE/-FO data, since TWS variability in the area is known to be very small. We would like to recall that return levels are estimated relative to trend and climatology, that is, after removing a linear trend and a (semi-)annual signal. To put the magnitudes of the extremes into perspective, we additionally compared them to the amplitude of the annual cycle (Figure S1 in Supporting Information S1). This comparison reveals that in many regions (especially in the northern hemisphere) the return levels are larger than the corresponding amplitudes of the seasonal signal and are, therefore, strongly noticeable in the time series. Only in regions with exceptionally strong annual signals, such as the Amazon region, the return levels are smaller than the amplitudes, but often still account for more than 50% of it (minimum: 38.4%).

As a first comparison to CMIP6 model data, we calculate 1-in-10-year return levels of TWS annual minima from a single experiment (r45i1p1f1) of the MPI-ESM1-2-LR model developed at the Max-Planck-Institute for Meteorology in Hamburg, Germany (Figure 2b). We note a generally similar spatial pattern of return level values when compared to GRACE/-FO, with notably small values in arid climates and highest values appearing in tropical and subtropical regions influenced by Monsoon dynamics. Interestingly, we also identify regionally coherent differences between model and observations, like substantially higher model-predicted return levels throughout Europe or much smaller values throughout Australia.

In order to utilize the full model-based information available from CMIP6, we also calculate return levels for the multi-model median (MMMed) of all 17 climate models listed above (Figure 2c). It is determined by calculating the return level for each CMIP simulation separately and subsequently building a weighted median of the estimated return levels. The weights are applied in such a way that each model has the same impact on the multi-model result, regardless of the number of individual simulations available. Let $x_{(i)}$ denote the i-th smallest return level from all multi-model ensemble members for a specific grid cell, and $w_{(i)}$ the corresponding weight assigned to this run. The MMMed is determined as $x_{(k)}$ such that

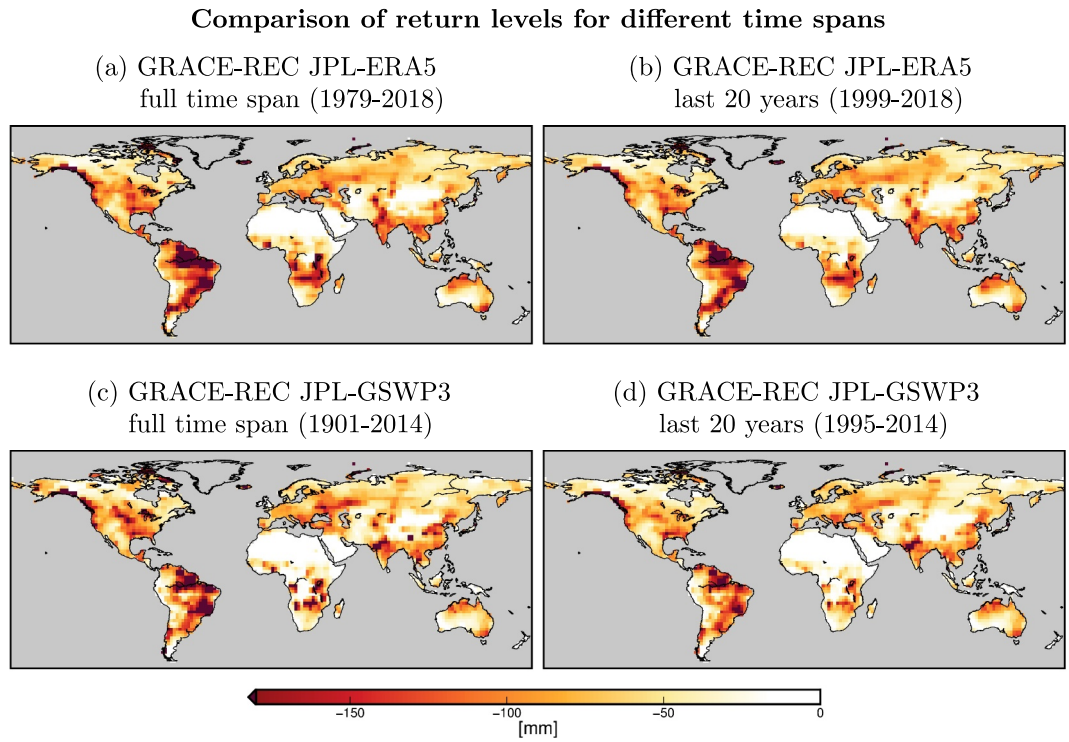

Figure 3. One-in-ten years return levels of TWS annual minima from the GRACE reconstructions based on ERA5 (top) and GSWP3 (bottom) for the full available time period (left) as well as for the most recent 20 years available from each data set (right).

$$\sum_{i=1}^{k} w_{(i)} \leq \frac{1}{2} \sum_{i=1}^{N} w_i < \sum_{i=1}^{k+1} w_{(i)} \tag{12}$$

applies. Here, $N = 245$ denotes the number of all simulations. As the median is calculated separately for each grid cell, the variability and range of values on a global level might not accurately reflect the range of the individual models. Therefore, we also calculate a scaled MMMed in Figure 2d by a method described in Jensen et al. (2020): For each individual model run, an empirical cumulative distribution function (ECDF) is calculated spatially and subsequently a multi-model median from all 245 ECDFs (mECDF) is derived. Additionally, the ECDF from the MMMed return level is calculated, which has a smaller range of values than the individual 245 ECDFs. Each value of the MMMed return level is then scaled to match the corresponding mECDF value at the same percentile as its ECDF. The scaled MMMed thus better reflects the spatial variability of the individual model results, whereas it may over- or under-estimate the model results analyzed in a single grid cell. It is therefore a useful complement to the unscaled MMMed for drawing conclusions about the general global pattern of the modeled return levels.

While return levels based on the unscaled MMMed are indeed much smaller than values observed by GRACE/-FO or predicted by MPI-ESM1-2-LR, we find the results from the scaled MMMed better in line with the results from the other two sources. This indicates that individual CMIP6 models do not predominantly underestimate the return levels derived by GRACE/-FO, as Figure 2c would suggest. Compared to Figure 2b, we find (scaled) MMMed return levels both exceeding (e.g., in Australia) and undercutting (e.g., Europe) the values obtained from MPI-ESM1-2-LR.

Since GRACE/-FO data is only available for 20 years, the results above might be influenced by the limited length of the time series. To quantify the impact, we revert back to GRACE-REC that is, available over 40 (JPL-ERA5) or even 114 years (JPL-GSWP3). In Figures 3a and 3c, the results are shown for the longest available data set of both reconstructions. Thus, these results are indicators for the differences in the return levels that would occur if

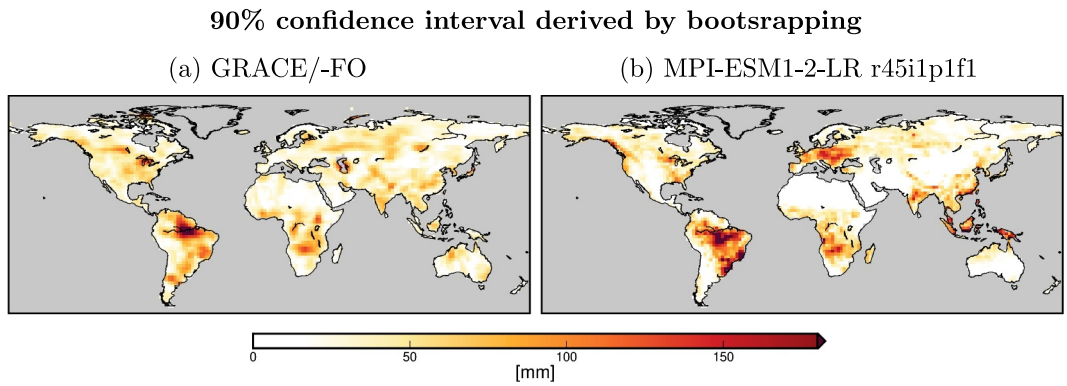

Figure 4. 90% confidence interval width for 1-in-10 years return levels of annual TWS minima for (a) the GRACE/-FO satellite data record and (b) a single realization (r49i1p1f1) of the MPI-ESM1-2-LR contribution to CMIP6 as obtained with a bootstrapping approach.

longer TWS observations from the past were available. In Figures 3b and 3d on the contrary, the return levels based on the last 20 years of the available time epoch are presented, where the same data gaps of 2 years for the annual minima are included as are present in GRACE/-FO. Throughout the world, return levels obtained from GRACE-REC are somewhat less extreme than those obtained from GRACE/-FO, indicating that the variability observed by the satellites exceeds those contained in the reconstructions. The return levels computed from both reconstructions for different time periods (i.e., full time span vs. the last 20 years), however, show very good agreement on the spatial patterns and also the magnitudes. The 114 years-long time series of GRACE-REC JPL-GSWP3 leads generally to slightly smaller, thus more extreme, return levels than the respective 20 years-long data set, whereas the difference between the 40 and the 20 years long time series of GRACE-REC JPL-ERA5 are minimal. It is therefore concluded that the observational basis currently available from GRACE/-FO is sufficient to calculate robust 1-in-10 years return levels for annual dry minima in TWS. Please note that the purpose of the reconstructions in our study is to justify the validity of our conclusions based on the limited GRACE/-FO record and not to serve as independent evaluation of the CMIP6 models.

5. Uncertainties of Return Levels for TWS Minima

We evaluate the uncertainties attached to the 1-in-10 years return levels in annual dry minima from CMIP6 model data and GRACE/-FO satellite observations by utilizing both parametric bootstrapping approaches and computationally much less demanding elementary ensemble statistics.

5.1. Parametric Bootstrapping

To perform parametric bootstrapping (e.g., Kyselý, 2008; Mudelsee, 2010), we use probability density functions characterized by the estimated GEV parameters to draw 1999 random samples, each with a sample size identical to the data-derived annual minima time series length. The number of samples was chosen according to Mudelsee (2010) after recommendation by Efron and Tibshirani (1993) to suppress simulation noise and to avoid interpolation for common percentage points. For each of these random samples, the GEV parameters are estimated, using the same fitting method as before, and the return levels are calculated subsequently. To generate the percentile 90% CI (Kyselý, 2008; Mudelsee, 2010), the 0.05 and 0.95 quantiles of the empirical distribution are selected and subsequently constitute the borders of the CI.

In Figure 4, the width of the respective CIs for return levels derived by GRACE/-FO and a single CMIP6 model experiment (r49i1p1f1 of MPI-ESM1-2-LR) are presented. The width describes the difference between the upper and lower limit of the CI. While in both cases the width of the CIs lies below 50% of the respective return levels for the majority of grid cells, in some regions, mainly where heavier-tailed distribution functions are estimated (i.e., shape >0), it reaches >70%. This is plausible, as a wider data range in the distribution leaves room for stronger deviations. It is also interesting to note that uncertainties estimated for MPI-ESM1-2-LR in Europe are substantially larger than the corresponding values obtained for GRACE/-FO, which suggests little confidence in the severity of the European droughts predicted by this particular climate model. We hypothesize that the

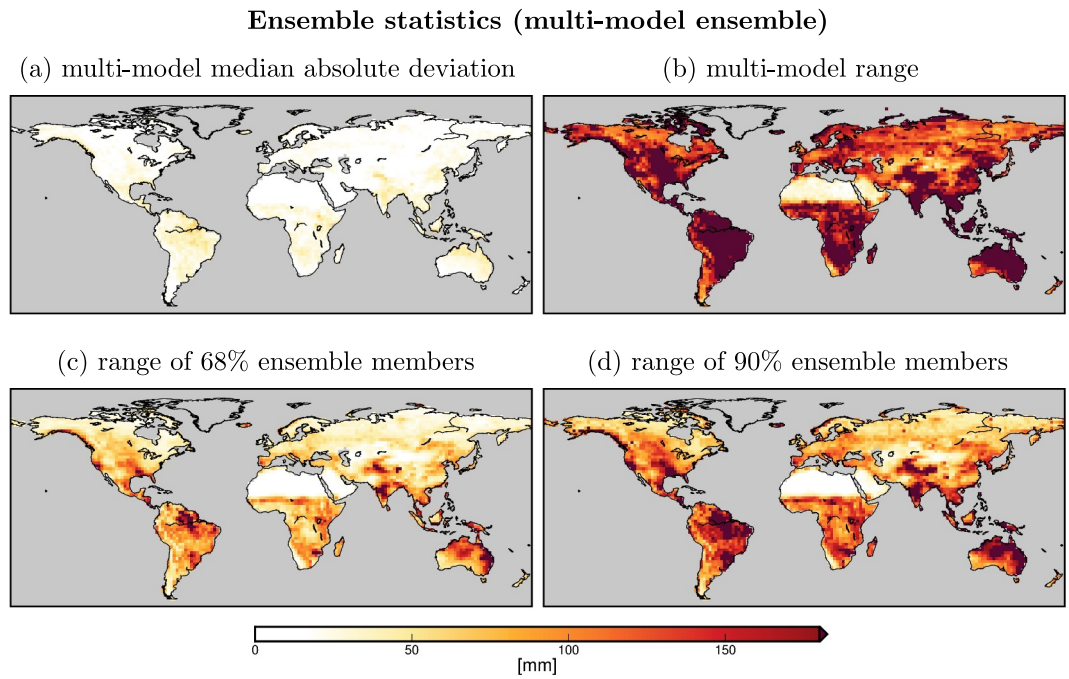

Figure 5. Characteristics of the multi-model CMIP6 ensemble for 1-in-10 years return levels of TWS annual minima: (a) multi-model median absolute deviation, (b) range of all ensemble members, (c) of 68% of all ensemble members and (d) of 90% of all ensemble members.

incorporation of additional observational data on interannual hydroclimate variability might help to narrow down this uncertainty in future versions of the MPI-ESM or its successors.

5.2. Uncertainty in Return Levels From the CMIP6 Multi-Model Ensemble

As a first measure of uncertainty in the ensemble-based 1-in-10 years return levels of annual TWS minima, we present the weighted multi-model Median Absolute Deviation (MAD) in Figure 5a. The MAD is a robust measure of variability (e.g., Hampel, 1974). To obtain it, we determine the absolute differences between each of the 245 individual model runs and the unscaled MMMed. We then calculate a weighted multi-model median using the method described earlier (see Equation 12). In most regions the multi-model MAD amounts to 15%–40% of the (unscaled) MMMed return levels. These rather small numbers results from the MAD being insensitive to individual outliers. Thus, it does not reflect how much individual models differ. Therefore, we additionally provide the multi-model range as the difference between the largest and lowest estimated return level from all model runs. To complement the picture, similar spreads are calculated for 68% (corresponding to twice the standard deviation in case of a normal distribution) and 90% of the ensemble members (Figures 5b–5d). The full model range (100th percentile) is sensitive to even the most extreme outliers in the model ensemble and can not be regarded as a representative uncertainty measure.

Since the range of 90% of the ensemble members is much higher than the bootstrapping-derived 90% CI from both GRACE/-FO and the MPI-ESM1-2-LR single run (Figure 4), which estimates the uncertainty associated with fitting the GEV distribution to the rather short data set, we conclude that for CMIP6-derived return levels the uncertainty resulting from the model formulation is much higher than that arising from the fitting process.

The same measures as in Figure 5 but only for the 50 members of the MPI-ESM1-2-LR model ensemble are shown in Figure 6. These individual runs only differ in their realization index, hence providing an impression for the internal variability of the individual model, which is much smaller than the variability of the multi-model ensemble. The range of 90% ensemble members has a similar magnitude as the fitting-related uncertainties estimated for the single r45i1p1f1 run of this model (cf. Figure 4b). Therefore, the uncertainty related to the internal variability in this model is in a similar range as the modeling uncertainty, and thus gives a good approximation for the latter. These results align qualitatively with Wehner (2010) who found that model

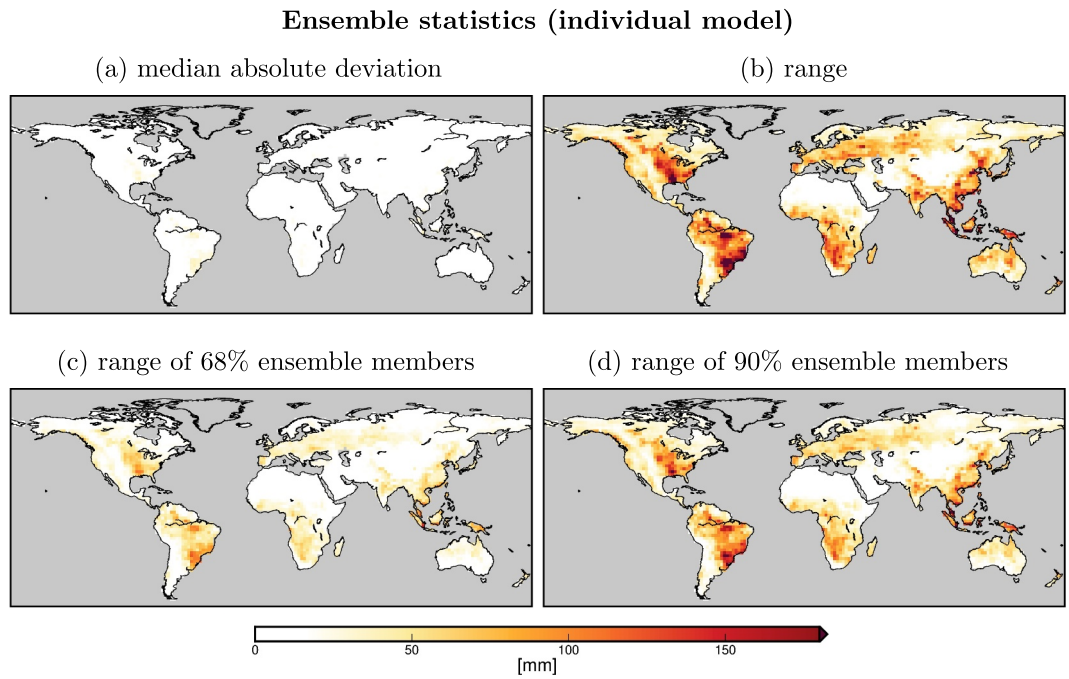

Figure 6. Characteristics of the MPI-ESM1-2-LR ensemble for 1-in-10 years return levels of TWS annual minima: (a) median absolute deviation, (b) range of all ensemble members, (c) of 68% of all ensemble members and (d) of 90% of all ensemble members.

formulation has a larger influence on the return levels of surface air temperature in (now outdated) CMIP3 climate models than the internal variability of the climate system itself and the uncertainty resulting from fitting GEV parameters.

6. Evaluation of Individual CMIP6 Models

A further aim is to identify how well return levels for the ensemble mean of each individual climate model fit to GRACE/-FO derived return levels. Since the uncertainty analysis has shown that the return level deviations from different models have a stronger impact on the multi-model result than the GEV fitting uncertainty, assessing these individual model results provides meaningful insights about the ability of the different climate models to simulate the occurrence of dry extremes in TWS variability. Two metrics are used to spatially evaluate the global fit of the ensemble mean of each model with GRACE/-FO, the Pearson correlation coefficient ρ and the root mean squared deviation (RMSD). While the first is a measure for the similarity of the global patterns, the latter takes into account the differences in magnitude. Both values are separately scaled between 0 and 1 with 0 standing in both cases for the worst fit (i.e., lowest correlation and highest RMSD) and 1 for the best fit (Figure 7a). The scaling thus reverses the order of the RMSD values in order to provide a consistent rating scale. Additionally, we compute the same quantities for the two GRACE-REC data sets, each evaluated for the longest available data record (Figures 7b and 7c).

It is clearly visible that some models perform better than others with respect to return levels derived from both GRACE/-FO and GRACE-REC. The reconstructions agree well with the model assessment by GRACE/-FO, suggesting that the results do not depend strongly on the observation time range (20 years vs. 40 or 114 years). A model that stands out in both spatial correlation coefficient and spatial RMSD, is the ACCESS-ESM1-5 model (compared with GRACE/-FO: $\rho = 0.62$ and RMSD = 38.0 mm). The model BCC-CSM2-MR shows a particularly bad performance regarding the correlation, with $\rho = 0.03$ compared to GRACE/-FO, whereas all other models have values between 0.3 and 0.6. The RMSD values compared to GRACE/-FO are in a range between 38 and 62 mm. The scaled and unscaled multi-model median yield the best results for both metrics in all three comparisons. This is in line with the usual assumption of ensemble statistics that the median from the results of a wide range of different models provides a better approximation of the true conditions than individual models. It is

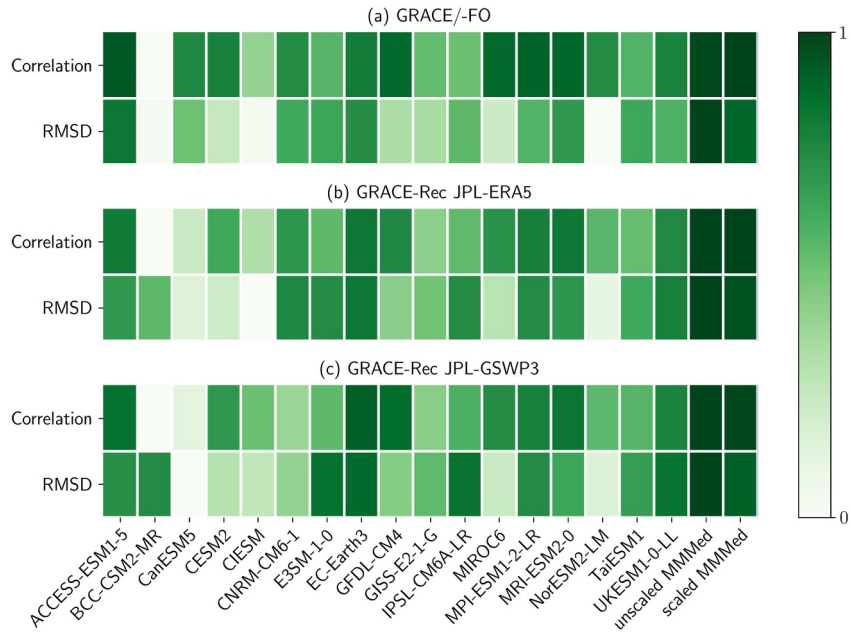

Figure 7. Root mean squared differences and correlations for single-model ensemble mean 1-in-10 years return levels of annual TWS minima for 17 individual model contributions to CMIP6 w.r.t. (a) GRACE/-FO satellite observations; (b) GRACE-like reconstruction based on ERA5; and (c) GRACE-like reconstruction based on GSWP3. All values are scaled between 0 and 1 with 1 representing the best fit. Additionally, the scaled and unscaled multi-model median results are provided.

worth emphasizing that it is also an indicator for the usability of GRACE/-FO for climate model validation and verification, that indeed the MMMed is identified as the best fit.

We recall that the mTWS changes are comprised of two storage compartments that are very differently represented in numerical climate models: soil moisture (`mrso`) and snow (`snw`). We repeat the analysis and compute correlation and RMSD for each of the two compartments by comparing the individual models (ensemble mean of each model) to the unscaled MMMed. The MMMed was chosen as reference since satellite observations of the individual compartments are not available to us. The results are shown in Supporting Information S1 (Figure S2). As expected, return levels in TWS are generally dominated by soil moisture estimates, which are available for all continental grid cells and not just in snow covered areas. Additionally, this individual assessment of `mrso` and `snw` gives insight into possible reasons for a good/bad fit shown in Figure 7. For example, the worst performing model (w.r.t. GRACE/-FO), which is BCC-CSM2-MR, does not agree well with the other models w.r.t. `mrso`, but shows a very good agreement for `snw`. The contrary is true for CanESM5: here `mrso` agrees well with the MMMed of all models, but the `snw` variable show large deviations. To identify particular reasons for those deviations will require collaboration with scientists responsible for individual experiments.

The return level results for a single simulation of the well-performing ACCESS-ESM1-5 model (r20i1p1f1) and the poorly fitting BCC-CSMS-MR simulation (r1i1p1f1) are displayed in Figures 8a and 8d, respectively. Strong deviations in the spatial pattern between the BCC-CSM2-MR and GRACE/-FO derived return levels, as indicated by the small correlation coefficient, are clearly visible, as return levels in the Global South are consistently underestimated by the model. The return level estimates from the ACCESS-ESM1-5 r20i1p1f1 run are similar in magnitude and pattern, as expected. Additionally to those two rather good and poor performing simulations, we also show examples that perform moderately well when compared to GRACE/-FO: one with particularly good overall correlation (Figure 8b) and the other with a good (i.e., small) RMSD (Figure 8c). In both cases, a similar pattern with different amplitudes or a different pattern with similar return levels is found, respectively, when compared to the satellite-derived return levels. Furthermore, the comparison of all four results of the individual example simulations points out some commonalities like regions of generally larger return levels, for example, in the Amazon area and Northern Australia (except for BCC-CSM2-MR). These commonalities agree mostly with the pattern of GRACE/-FO derived return levels. Nevertheless, Figure 8 also highlights the rather large

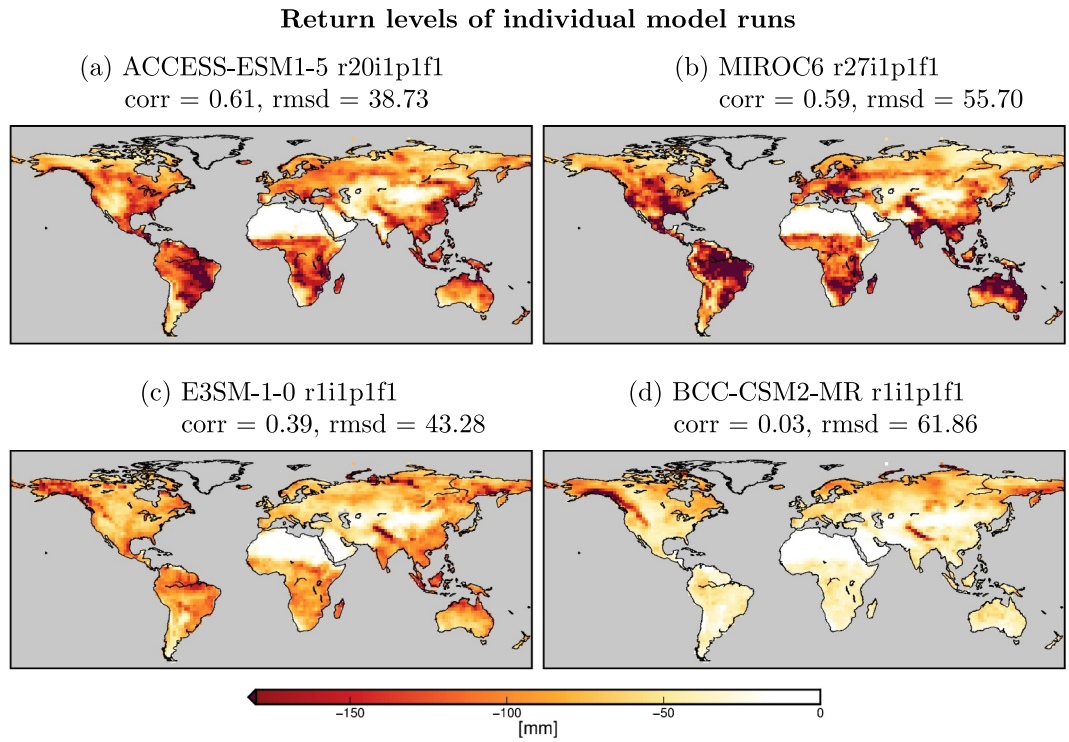

Figure 8. One-in-ten years return levels of TWS annual minima from 2002 to 2022 for individual contributions to CMIP6: (a) the overall best fitting experiment r20i1p1f1 from ACCESS-ESM1-5; (b) high correlation for experiment r27i1p1f1 from MIROC6; (c) small RMSD for experiment r1i1p1f1 from E3SM-1-0; and (d) a rather poorly performing experiment r1i1p1f1 from BCC-CSM2-MR.

differences in the magnitude and more detailed spatial pattern of the return levels of annual minima and therefore the need to evaluate climate models regarding mTWS. For all 17 models the spatial patterns of the return levels for mTWS (Figure S3 in Supporting Information S1) and for the two compartments individually (Figures S4 and S5) are shown in Supporting Information S1.

7. Long-Term Changes of One-in-Ten Years Return Levels

7.1. Emission Scenario SSP5-8.5

In order to examine the evolution of dry extremes predicted by the climate models, we compare the return levels derived from 30-year epochs at the beginning and the end of the 21st century by calculating their differences. In Figure 9a the (unscaled) MMMed of these differences are shown. Here, negative values correspond to an increase in the intensity of extreme values. To see how this compares to the previous century, in Figure 9c the respective differences for the 20th century are presented. Additionally, it is investigated how many models share the same tendency to project more or less extreme return levels by counting how many models agree on the sign of the differences in the ensemble means of their return levels. If 75%, that is, 13 out of 17 models, agree on the sign, the regions are considered as high consensus regions (Figures 9b and 9d).

The difference in the return levels between present and future conditions (Figure 9a) is dominated by a predicted increase in the deficits (i.e., yellow to red colors). This is particularly the case in regions with strong expected changes amounting to more than, e.g., 0.1 mm/yr change visible in the MMMed, such as central Africa, South-East-Asia, the western Amazonas region, and Northern Siberia. Overall, an increase in the return level magnitude is predicted for 70% of the land area considered. A decrease (i.e., a tendency toward less extreme values) is visible in some regions as well, for example, South of Africa, Gulf Coast of Alaska, Northern Scandinavia and the lower Amazon catchment. The regions where the most pronounced changes are visible, are in most cases also the regions with high model consensus, which further supports the predicted changes in these areas. In low consensus areas, often smaller changes are predicted by the multi-model differences, which can have two causes, either

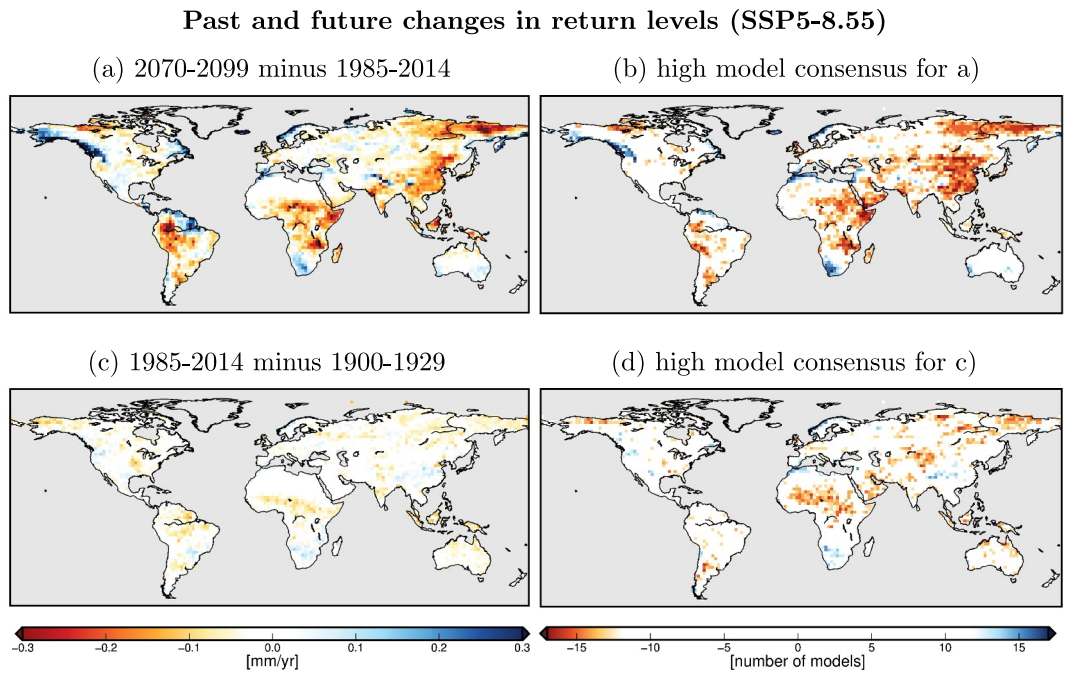

Figure 9. Change in 1-in-10 years return levels of TWS annual minima as seen by CMIP6 (left) between present day and future conditions at the end of the 21st century under the SSP5-8.5 scenario (top) and between early industrial times and present day (bottom) together with a depiction of the regions of high consensus among the different models participating in CMIP6 (right).

smaller predicted changes by the majority of the models or almost equally strong trends in both directions that cancel out.

Compared to the changes during the 20th century (Figure 9c), there is a strong increase in the magnitude of the predicted changes for the 21st century. This is true for both directions, that is, becoming more or less extreme. The models also disagree more on the direction of change for the earlier century (Figure 9b), with only a few high consensus areas, mainly in central Africa and smaller parts of Asia. This leads us to conclude that the climate models predict much stronger and more consistent changes in the return levels of TWS minima over the 21st century compared to the 20th century.

It is important to note that the predicted changes for a certain region do not necessarily reflect in an increase or decrease of absolute dry extremes of TWS, as the removed linear trend and seasonal cycles can possibly counteract these tendencies. As discussed above, the return levels that were investigated in this study rather correlate with the temporal variability of the data. Therefore, the return levels of the minima can be expected to evolve similarly as the variability of the data. To examine this, we compare our results with Figure 9 of Jensen et al. (2020), showing the change in the interannual RMS of TWS in CMIP6 over the period 2000–2100. Predicted changes in the return levels derived in this study correspond closely to the predicted change in the innterannual RMS, thereby confirming that increased atmospheric temperatures—and consequently atmospheric water-holding capacities as quantified by the Clausius-Clapeyron equation—lead to an intensification of the global hydrological cycle (Huntington, 2006). This intensification is also accompanied by the increasing severity of hydrometeorological extremes including both wet and in particular also dry events as demonstrated in this work.

7.2. Comparison to Other Emission Scenarios

In the section above, expected changes in return levels (Figure 9a) were analyzed for the most severe high-emissions pathway SSP5-8.5 included in CMIP6. It assumes limited climate policies, resulting in high greenhouse gas emissions and a radiative forcing of 8.5 W/m^2 by the year 2100 (O'Neill et al., 2016). While this might very well be a realistic assumption, in the following the expected changes in return levels from the more optimistic scenarios SSP1-2.6, SSP2-4.5, SSP3-7.0 are analyzed for comparison. Figure 10 clearly shows the tendency toward

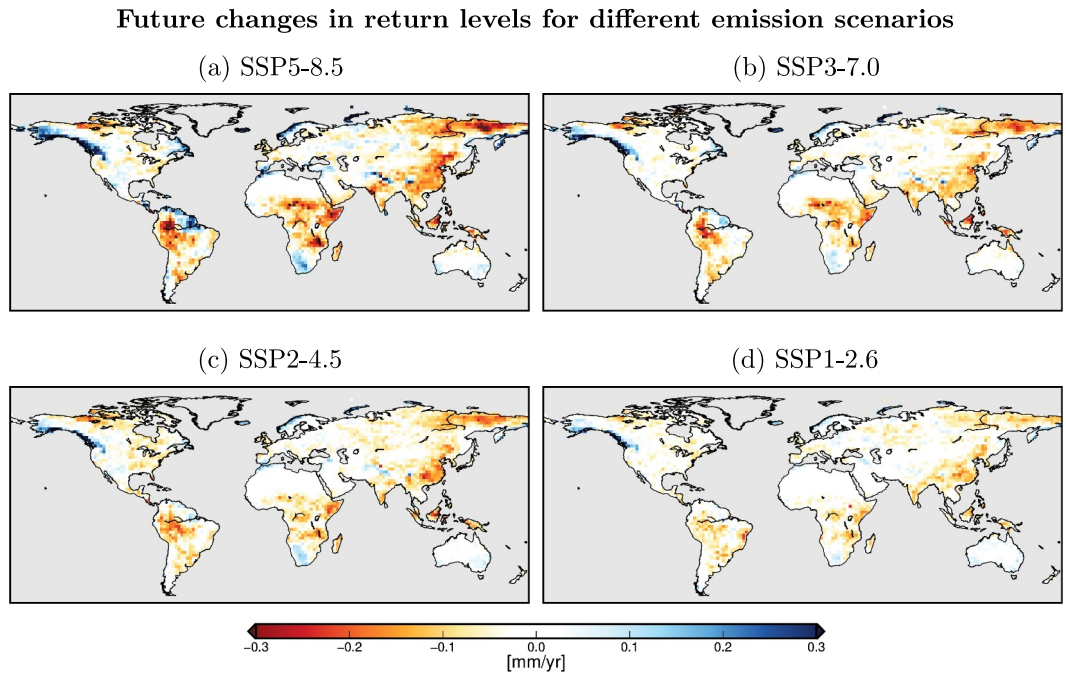

Figure 10. Change in 1-in-10 years return levels of TWS annual minima as seen by CMIP6 between present day and future conditions at the end of the 21st century under the SSP5-8.5 scenario (a, same as Figure 9a, 17 models with 245 model runs) in comparison to the same values for SSP3-7.0 (b, 15 models with 280 model runs), SSP2-4.5 (c, 16 models with 260 model runs), and SSP1-2.6 (d, 15 models with 248 model runs).

more extreme values (red colors) for stronger greenhouse gas emission pathways. The maximum change in return levels toward more dry extremes projected by the SSP5-8.5 pathway (Figure 10a, same as Figure 9a) amounts to −0.42 mm/yr (median: −0.031 mm/yr) compared to smaller maximum rates of change of −0.31 mm/yr for SSP3-7.0, (Figure 10b, median: 0.021 mm/yr), and −0.24 mm/yr for SSP1-2.6 (Figure 10d, median: −0.017 mm/yr). We recall that these numbers refer to the unscaled MMMed and the expected rate-of-change simulated by individual models is considerably larger. Please note that not all models participating in CMIP6 provide simulations for each of the four scenarios and if they do, the number of simulations per scenario often differs slightly. In our comparison, we focus on the 17 selected models that provide results for the SSP5-8.5 scenario (see Table 1), as these are the models that have been used primarily in this manuscript. Since not all of these 17 models also deliver results for the other scenarios, slightly different numbers of model runs (see caption of Figure 10) prevent the results from being perfectly comparable. Despite those technical limitations, the results clearly confirm the expected increase in extreme events with increasing greenhouse gas emissions demonstrating that there is a notable difference in those indicators of the hydroclimate for different socioeconomic pathways that humanity is still able to choose.

8. Summary and Conclusions

From a 20 years-long satellite record of terrestrial water storage variations, 1-in-10 years return levels of TWS minima have been calculated by means of the generalized extreme value (GEV) distribution and subsequently compared against numerical coupled climate model data from the CMIP6 archive. Both spatial patterns and also magnitudes of the return levels for extreme dry events agree reasonably well in models and observations, with the largest deviations generally found in areas characterized by semi-humid climate conditions. Among all individual model data sets tested, GRACE/-FO observations identify the multi-model median (MMMed) as the best-fitting representation of the 1-in-10 years return levels of TWS minima, thereby well confirming the inherent assumptions of multi-model ensemble assessments which expect the MMMed to be the best possible approximation of the truth. The fact that GRACE/-FO can indeed identify the MMMed among all tested realizations adds further credibility to the satellite observations as a novel tool for assessing climate model data. The magnitude of the return levels derived from the satellite observations relative to the expected long-term deterministic signal

(characterized by linear trend and the (semi-)annual cycle) clearly exceeds the magnitude of seasonal changes (i.e., the amplitude of the annual cycle) and will, therefore, be clearly identifiable in the water storage time series.

The uncertainties of return levels were quantified using a parametrized bootstrapping algorithm; derived uncertainties have been subsequently connected to more frequently utilized error metrics based on ensemble statistics. It was seen that the derived return levels from GRACE-/FO and a single CMIP6 model run (MPI-ESM1-2-LR) come with not negligible uncertainties that are inherent from the short observation time period of 20 years. Nevertheless, these (fitting) uncertainties are smaller than the differences between return levels from several CMIP6 models. Thus, we were able to identify differences in the level of agreement of individual climate models with respect to the observation-based GRACE/-FO derived return levels. We note, however, that temporal autocorrelation within the TWS time series might still persist, since any past dry event might deplete the deep water storages to an extent that is, not rechargeable by average rainfall amounts. We address this problem by using only a single value from each year (annual minima) for the estimation of the 1-in-10-year return levels. However, prolonged multi-year drought events, will enter the estimation as multiple consecutive annual minima, violating the assumption of independent temporal blocks as an essential prerequisite for the application of the GEV theory. Please note that broader blocks are not possible due to short observational time series of just 20 years. It also needs to be considered that the data might not only be non-stationary, but also not identically distributed given the strong changes in the Earth's hydrometeorological conditions induced by the steeply increasing greenhouse gas concentrations. CMIP6 models indicate that long-term changes in return levels from the beginning of the 20th century until today have been modest so far, whereas much stronger changes are predicted toward the end of the current century from the SSP5-8.5 scenario pathway. A comparison with other scenarios clearly reveals an increase in the severity of extreme events with increasing greenhouse gas emissions demonstrating that there is a notable difference in the evolution of the hydroclimate for different possible socioeconomic pathways.

For future studies, the fact that return levels in individual models deviate strongly from the observations offers opportunities to inform, validate and/or calibrate numerical models with satellite gravimetry data in order to further improve model skill particularly in regions affected by monsoon climates. Another objective for future studies would be the assessment of longer return horizons (e.g., 1-in-30 years, 1-in-50 years, 1-in-100 years) required for infrastructure planning decisions. While nowadays such assessments are still associated with high uncertainties, the growing observational record will make such analyses increasingly more reliable. Similarly to the characteristics of dry events considered in this paper, the elaborated workflow could be also applied to wet extremes, which might be slightly more challenging due to the usually short duration of floods that are not necessarily well captured by the nominally monthly sampling of the GRACE/-FO products. It might be thus necessary to also consider dedicated low-resolution gravity time series with daily sampling obtained with a Kalman Smoother approach provided along-side the nominal ITSG-Grace2018 monthly solutions used throughout this paper to study hydrometeorological wet extremes.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

CMIP6 data are publicly available from the Earth System Grid Federation (ESGF) Data Portal (ESGF CMIP6, 2019), see also Eyring et al. (2016). The CMIP6 models used in this study are listed in Table 1. ITSG-Grace2018 monthly mean gravity fields are accessible via Mayer-Gürr et al. (2018) or Kvas et al. (2019). The GRACE-REC data sets are provided by Humphrey and Gudmundsson (2019a, 2019b).

Acknowledgments
This study has been funded by the German Research Foundation (DFG) via the Research Unit NEROGRAV (FOR 2736) under the Grant EI 772/3-2. LJ has been funded by the DFG under Grant SA 2952/5-1. We gratefully acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modeling, coordinated and promoted CMIP6. We also thank the climate modeling groups for producing and making available their model output, the ESGF for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and ESGF. Open Access funding enabled and organized by Projekt DEAL.

References

Bader, D. C., Leung, R., Taylor, M., & McCoy, R. B. (2019). E3SM-project E3SM1.0 model output prepared for CMIP6 CMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.2294

Bader, D. C., Leung, R., Taylor, M., & McCoy, R. B. (2022). E3SM-project E3SM1.0 model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.15102

Berg, A., Sheffield, J., & Milly, P. C. D. (2017). Divergent surface and total soil moisture projections under global warming. *Geophysical Research Letters*, *44*(1), 236–244. https://doi.org/10.1002/2016GL071921

Boucher, O., Denvil, S., Levavasseur, G., Cozic, A., Caubel, A., Foujols, M.-A., et al. (2018). IPSL IPSL-CM6A-LR model output prepared for CMIP6 CMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.1534

Boucher, O., Denvil, S., Levavasseur, G., Cozic, A., Caubel, A., Foujols, M.-A., et al. (2019). IPSL IPSL-CM6A-LR model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.1532

Carlson, G., Werth, S., & Shirzaei, M. (2024). A novel hybrid GNSS, GRACE, and InSAR joint inversion approach to constrain water loss during a record-setting drought in California. *Remote Sensing of Environment*, *311*(7), 114303. https://doi.org/10.1016/j.rse.2024.114303

Carvalho, D., Rafael, S., Monteiro, A., Rodrigues, V., Lopes, M., & Rocha, A. (2022). How well have CMIP3, CMIP5 and CMIP6 future climate projections portrayed the recently observed warming. *Scientific Reports*, *12*(1), 11983. https://doi.org/10.1038/s41598-022-16264-6

Castillo, E., Hadi, A. S., Balakrishnan, N., & Sarabia, J. M. (2005). *Extreme value and related models with applications in engineering and science*. Wiley.

Chao, B. F., & Liau, J. R. (2019). Gravity changes due to large earthquakes detected in GRACE satellite data via empirical orthogonal function analysis. *Journal of Geophysical Research: Solid Earth*, *124*(3), 3024–3035. https://doi.org/10.1029/2018JB016862

Chen, Z., Zhou, T., Zhang, L., Chen, X., Zhang, W., & Jiang, J. (2020). Global land monsoon precipitation changes in CMIP6 projections. *Geophysical Research Letters*, *47*(14), 1225. https://doi.org/10.1029/2019GL086902

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

Chikobvu, D., & Chifurira, R. (2015). Modelling of extreme minimum rainfall using generalised extreme value distribution for Zimbabwe. *South African Journal of Science*, *111*(9/10), 8. https://doi.org/10.17159/SAJS.2015/20140271

Chou, C., Neelin, J. D., Chen, C.-A., & Tu, J.-Y. (2009). Evaluating the "Rich-Get-Richer" mechanism in tropical precipitation change under global warming. *Journal of Climate*, *22*(8), 1982–2005. https://doi.org/10.1175/2008JCLI2471.1

Coles, S. (2001). *An introduction to statistical modeling of extreme values*. Springer. https://doi.org/10.1007/978-1-4471-3675-0

Danabasoglu, G. (2019a). NCAR CESM2 model output prepared for CMIP6 CMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.2185

Danabasoglu, G. (2019b). NCAR CESM2 model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.2201

Davison, A. C., & Huser, R. (2015). Statistics of extremes. *Annual Review of Statistics and Its Application*, *2*(1), 203–235. https://doi.org/10.1146/annurev-statistics-010814-020133

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

EC-Earth Consortium. (2019a). EC-Earth-Consortium EC-Earth3 model output prepared for CMIP6 CMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.181

EC-Earth Consortium. (2019b). EC-Earth-Consortium EC-Earth3 model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.251

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Eicker, A., Schawohl, L., Middendorf, K., Bagge, M., Jensen, L., & Dobslaw, H. (2024). Influence of GIA uncertainty on climate model evaluation with GRACE/GRACE–FO satellite gravimetry data. *Journal of Geophysical Research: Solid Earth*, *129*(5), e2023JB027769. https://doi.org/10.1029/2023JB027769

ESGF CMIP6. (2019). CMIP6—Coupled model intercomparison project phase 6 [Dataset]. *World Climate Research Programme (WCRP)*. Retrieved from https://aims2.llnl.gov/search/cmip6/

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

Fisher, R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, *24*(2), 180–190. https://doi.org/10.1017/S0305004100015681

Flechtner, F., Wiese, D., Webb, F., Landerer, F., Gross, M., Snopek, K., et al. (2024). Global gravity and mass change observations beyond GRACE-FO: Updates on the upcoming GRACE-Continuity mission. In *GRACE/GRACE-FO Science Team Meeting*. GSTM2024-21. https://doi.org/10.5194/gstm2024-21

Frich, P., Alexander, L. V., Della-Marta, P., Gleason, B., Haylock, M., Klein Tank, A. M., & Peterson, T. (2002). Observed coherent changes in climatic extremes during the second half of the twentieth century. *Climate Research*, *19*, 193–212. https://doi.org/10.3354/cr019193

Früh, B., Feldmann, H., Panitz, H.-J., Schädler, G., Jacob, D., Lorenz, P., & Keuler, K. (2010). Determination of precipitation return values in complex terrain and their evaluation. *Journal of Climate*, *23*(9), 2257–2274. https://doi.org/10.1175/2009JCLI2685.1

García-Cueto, O. R., Cavazos, M. T., de Grau, P., & Santillán-Soto, N. (2014). Analysis and modeling of extreme temperatures in several cities in northwestern Mexico under climate change conditions. *Theoretical and Applied Climatology*, *116*(1–2), 211–225. https://doi.org/10.1007/s00704-013-0933-x

Gerdener, H., Engels, O., & Kusche, J. (2020). A framework for deriving drought indicators from the Gravity Recovery and Climate Experiment (GRACE). *Hydrology and Earth System Sciences*, *24*(1), 227–248. https://doi.org/10.5194/hess-24-227-2020

Gnedenko, B. (1943). Sur La Distribution Limite Du Terme Maximum D'Une Serie Aleatoire. *Annals of Mathematics*, *44*(3), 423. https://doi.org/10.2307/1968974

Good, P., Sellar, A., Tang, Y., Rumbold, S., Ellis, R., Kelley, D., et al. (2019). MOHC UKESM1.0-LL model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.1567

Gubareva, T. S., & Gartsman, B. I. (2010). Estimating distribution parameters of extreme hydrometeorological characteristics by L-moments method. *Water Resources*, *37*(4), 437–445. https://doi.org/10.1134/S0097807810040020

Guo, H., John, J. G., Blanton, C., McHugh, C., Nikonov, S., Radhakrishnan, A., et al. (2018a). NOAA-GFDL GFDL-CM4 model output [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.1402

Guo, H., John, J. G., Blanton, C., McHugh, C., Nikonov, S., Radhakrishnan, A., et al. (2018b). NOAA-GFDL GFDL-CM4 model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.9242

Hamdi, Y., Duluc, C.-M., & Rebour, V. (2018). Temperature extremes: Estimation of non-stationary return levels and associated uncertainties. *Atmosphere*, *9*(4), 129. https://doi.org/10.3390/atmos9040129

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, *69*(346), 383. https://doi.org/10.2307/2285666

Hardy, R. A., Nerem, R. S., & Wiese, D. N. (2017). The impact of atmospheric modeling errors on GRACE estimates of mass loss in Greenland and Antarctica. *Journal of Geophysical Research: Solid Earth*, *122*(12), 10440–10458. https://doi.org/10.1002/2017JB014556

Held, I. M., & Soden, B. J. (2006). Robust responses of the hydrological cycle to global warming. *Journal of Climate*, *19*(21), 5686–5699. https://doi.org/10.1175/JCLI3990.1

Horvath, A., Murböck, M., Pail, R., & Horwath, M. (2018). Decorrelation of GRACE time variable gravity field solutions using full covariance information. *Geosciences*, *8*(9), 323. https://doi.org/10.3390/geosciences8090323

Huang, W. (2019a). THU CIESM model output prepared for CMIP6 CMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.1352

Huang, W. (2019b). THU CIESM model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.1357

Huang, X., Wang, Y., & Ma, X. (2024). Simulation of extreme precipitation changes in Central Asia using CMIP6 under different climate scenarios. *Theoretical and Applied Climatology*, *155*(4), 3203–3219. https://doi.org/10.1007/s00704-023-04802-9

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

Humphrey, V., & Gudmundsson, L. (2019b). GRACE-REC: A reconstruction of climate-driven water storage changes over the last century [Dataset]. *figshare*. https://doi.org/10.6084/m9.figshare.7670849.v3

Huntington, T. G. (2006). Evidence for intensification of the global water cycle: Review and synthesis. *Journal of Hydrology*, *319*(1–4), 83–95. https://doi.org/10.1016/j.jhydrol.2005.07.003

Jensen, L., Eicker, A., Dobslaw, H., & Pail, R. (2020). Emerging changes in terrestrial water storage variability as a target for future satellite gravity missions. *Remote Sensing*, *12*(23), 3898. https://doi.org/10.3390/rs12233898

Jensen, L., Eicker, A., Dobslaw, H., Stacke, T., & Humphrey, V. (2019). Long–term wetting and drying trends in land water storage derived from GRACE and CMIP5 models. *Journal of Geophysical Research: Atmospheres*, *124*(17–18), 9808–9823. https://doi.org/10.1029/2018JD029989

Katz, R. W., Parlange, M. B., & Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources*, *25*(8–12), 1287–1304. https://doi.org/10.1016/S0309-1708(02)00056-8

Kochanek, K., Renard, B., Arnaud, P., Aubert, Y., Lang, M., Cipriani, T., & Sauquet, E. (2014). A data-based comparison of flood frequency analysis methods used in France. *Natural Hazards and Earth System Sciences*, *14*(2), 295–308. https://doi.org/10.5194/nhess-14-295-2014

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

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

Kvas, A., Behzadpour, S., Ellmer, M., Klinger, B., Strasser, S., Zehentner, N., & Mayer-Gürr, T. (2019). ITSG–Grace2018: Overview and evaluation of a new GRACE–Only gravity field time series. *Journal of Geophysical Research: Solid Earth*, *124*(8), 9332–9344. https://doi.org/10.1029/2019JB017415

Kyselý, J. (2008). A cautionary note on the use of nonparametric bootstrap for estimating uncertainties in extreme-value models. *Journal of Applied Meteorology and Climatology*, *47*(12), 3236–3251. https://doi.org/10.1175/2008JAMC1763.1

Lambeck, K. (1988). *Geophysical geodesy: The slow deformations of the Earth*. Clarendon Press.

Landerer, F. W., Flechtner, F. M., Save, H., Webb, F. H., Bandikova, T., Bertiger, W. I., et al. (2020). Extending the global mass change data record: GRACE follow–On instrument and science data performance. *Geophysical Research Letters*, *47*(12), Si. https://doi.org/10.1029/2020GL088306

Leadbetter, M. R., Lindgren, G., & Rootzén, H. (1983). *Extremes and related properties of random sequences and processes*. Springer New York. https://doi.org/10.1007/978-1-4612-5449-2

Lee, W.-L., & Liang, H.-C. (2019). AS-RCEC TaiESM1.0 model output prepared for CMIP6 CMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.9684

Lee, W.-L., & Liang, H.-C. (2020). AS-RCEC TaiESM1.0 model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.9688

Li, L., Li, P., & Liu, Y. (2014). How we determine the design environmental conditions and how they impact the structural reliabilities? In *Structures, safety and reliability* (Vol. 4A). American Society of Mechanical Engineers. https://doi.org/10.1115/OMAE2014-23198

Martins, E. S., & Stedinger, J. R. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, *36*(3), 737–744. https://doi.org/10.1029/1999WR900330

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

Moon, S., & Ha, K.-J. (2020). Future changes in monsoon duration and precipitation using CMIP6. *Npj Climate and Atmospheric Science*, *3*(1), 711. https://doi.org/10.1038/s41612-020-00151-w

Morrison, J. E., & Smith, J. A. (2002). Stochastic modeling of flood peaks using the generalized extreme value distribution. *Water Resources Research*, *38*(12), 1. https://doi.org/10.1029/2001WR000502

Mudelsee, M. (2010). *Climate time series analysis: Classical statistical and bootstrap methods* (Vol. 42). Springer.

NASA Goddard Institute for Space Studies (NASA/GISS). (2018). NASA-GISS GISS-E2.1G model output prepared for CMIP6 CMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.1400

NASA Goddard Institute for Space Studies (NASA/GISS). (2020). NASA-GISS GISS-E2.1G model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.2074

Nerantzaki, S. D., & Papalexiou, S. M. (2022). Assessing extremes in hydroclimatology: A review on probabilistic methods. *Journal of Hydrology*, *605*(68), 127302. https://doi.org/10.1016/j.jhydrol.2021.127302

O'Neill, B. C., Tebaldi, C., van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., et al. (2016). The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6. *Geoscientific Model Development*, *9*(9), 3461–3482. https://doi.org/10.5194/gmd-9-3461-2016

Panda, D. K., & Wahr, J. (2016). Spatiotemporal evolution of water storage changes in India from the updated GRACE –derived gravity records. *Water Resources Research*, *52*(1), 135–149. https://doi.org/10.1002/2015WR017797

Panet, I., Narteau, C., Lemoine, J.-M., Bonvalot, S., & Remy, D. (2022). Detecting preseismic signals in GRACE gravity solutions: Application to the 2011 Tohoku M w 9.0 earthquake. *Journal of Geophysical Research: Solid Earth*, *127*(8), 28. https://doi.org/10.1029/2022JB024542

Papalexiou, S. M., & Koutsoyiannis, D. (2013). Battle of extreme value distributions: A global survey on extreme daily rainfall. *Water Resources Research*, *49*(1), 187–201. https://doi.org/10.1029/2012WR012557

Peltier, W. R., Argus, D. F., & Drummond, R. (2018). Comment on "An assessment of the ICE–6G_C (VM5a) glacial isostatic adjustment model" by Purcell et al. *Journal of Geophysical Research: Solid Earth*, *123*(2), 2019–2028. https://doi.org/10.1002/2016JB013844

Reager, J. T., & Famiglietti, J. S. (2009). Global terrestrial water storage capacity and flood potential using GRACE. *Geophysical Research Letters*, *36*(23), H1. https://doi.org/10.1029/2009GL040826

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

Reiss, R.-D., & Thomas, M. (2007). *Statistical analysis of extreme values: With applications to insurance, finance, hydrology and other fields* (3rd ed.). Birkhäuser Verlag AG. https://doi.org/10.1007/978-3-7643-7399-3

Rodell, M., Beaudoing, H. K., L'Ecuyer, T. S., Olson, W. S., Famiglietti, J. S., Houser, P. R., et al. (2015). The observed state of the water cycle in the early twenty-first century. *Journal of Climate*, *28*(21), 8289–8318. https://doi.org/10.1175/JCLI-D-14-00555.1

Scanlon, B. R., Longuevergne, L., & Long, D. (2012). Ground referencing GRACE satellite estimates of groundwater storage changes in the California Central Valley, USA. *Water Resources Research*, *48*(4), 116. https://doi.org/10.1029/2011WR011312

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

Seland, Ø., Bentsen, M., Oliviè, D. J. L., Toniazzo, T., Gjermundsen, A., Graff, L. S., et al. (2019a). NCC NorESM2-LM model output prepared for CMIP6 CMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.502

Seland, Ø., Bentsen, M., Oliviè, D. J. L., Toniazzo, T., Gjermundsen, A., Graff, L. S., et al. (2019b). NCC NorESM2-LM model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.604

Shim, S., Lim, Y.-J., Byun, Y.-H., Seo, J., Kwon, S., & Kim, B.-H. (2020a). NIMS-KMA UKESM1.0-LL model output prepared for CMIP6 CMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.2245

Shim, S., Lim, Y.-J., Byun, Y.-H., Seo, J., Kwon, S., & Kim, B.-H. (2020b). NIMS-KMA UKESM1.0-LL model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.2250

Shiogama, H., Abe, M., & Tatebe, H. (2019). MIROC MIROC6 model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.898

Stephenson, A., & Tawn, J. (2004). Bayesian inference for extremes: Accounting for the three extremal types. *Extremes*, *7*(4), 291–307. https://doi.org/10.1007/s10687-004-3479-6

Stevenson, S., Huang, X., Zhao, Y., Di Lorenzo, E., Newman, M., Xu, T., & Capotondi, A. (2023a). UCSB E3SM1.0 model output prepared for CMIP6 CMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.17062

Stevenson, S., Huang, X., Zhao, Y., Di Lorenzo, E., Newman, M., Xu, T., & Capotondi, A. (2023b). UCSB E3SM1.0 model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.17065

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

Swart, N. C., Cole, J. N., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., et al. (2019a). CCCma CanESM5 model output prepared for CMIP6 CMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.1303

Swart, N. C., Cole, J. N., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., et al. (2019b). CCCma CanESM5 model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.1317

Swenson, S., Chambers, D., & Wahr, J. (2008). Estimating geocenter variations from a combination of GRACE and ocean model output. *Journal of Geophysical Research*, *113*(B8), 29077. https://doi.org/10.1029/2007JB005338

Swenson, S., & Wahr, J. (2006). Post–processing removal of correlated errors in GRACE data. *Geophysical Research Letters*, *33*(8), L08402. https://doi.org/10.1029/2005GL025285

Tang, Y., Rumbold, S., Ellis, R., Kelley, D., Mulcahy, J., Sellar, A., et al. (2019). MOHC UKESM1.0-LL model output prepared for CMIP6 CMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.1569

Tapley, B. D., Watkins, M. M., Flechtner, F., Reigber, C., Bettadpur, S., Rodell, M., et al. (2019). Contributions of GRACE to understanding climate change. *Nature Climate Change*, *5*(5), 358–369. https://doi.org/10.1038/s41558-019-0456-2

Tatebe, H., & Watanabe, M. (2018). MIROC MIROC6 model output prepared for CMIP6 CMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.881

Tiwari, V. M., Wahr, J., & Swenson, S. (2009). Dwindling groundwater resources in northern India, from satellite gravity observations. *Geophysical Research Letters*, *36*(18), 321. https://doi.org/10.1029/2009GL039401

Vissa, N. K., Anandh, P. C., Behera, M. M., & Mishra, S. (2019). ENSO-induced groundwater changes in India derived from GRACE and GLDAS. *Journal of Earth System Science*, *128*(5), 1001. https://doi.org/10.1007/s12040-019-1148-z

Voldoire, A. (2018). CNRM-CERFACS CNRM-CM6-1 model output prepared for CMIP6 CMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.1375

Voldoire, A. (2019). CNRM-CERFACS CNRM-CM6-1 model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.1384

Wahr, J., Molenaar, M., & Bryan, F. (1998). Time variability of the Earth's gravity field: Hydrological and oceanic effects and their possible detection using GRACE. *Journal of Geophysical Research*, *103*(B12), 30205–30229. https://doi.org/10.1029/98JB02844

Wang, C.-H., & Holmes, J. D. (2020). Exceedance rate, exceedance probability, and the duality of GEV and GPD for extreme hazard analysis. *Natural Hazards*, *102*(3), 1305–1321. https://doi.org/10.1007/s11069-020-03968-z

Wang, J., Han, Y., Stein, M. L., Kotamarthi, V. R., & Huang, W. K. (2016). Evaluation of dynamically downscaled extreme temperature using a spatially-aggregated generalized extreme value (GEV) model. *Climate Dynamics*, *47*(9–10), 2833–2849. https://doi.org/10.1007/s00382-016-3000-3

Wehner, M. (2010). Sources of uncertainty in the extreme value statistics of climate data. *Extremes*, *13*(2), 205–217. https://doi.org/10.1007/s10687-010-0105-7

Wieners, K.-H., Giorgetta, M., Jungclaus, J., Reick, C., Esch, M., Bittner, M., et al. (2019a). MPI-M MPIESM1.2-LR model output prepared for CMIP6 CMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.742

Wieners, K.-H., Giorgetta, M., Jungclaus, J., Reick, C., Esch, M., Bittner, M., et al. (2019b). MPI-M MPIESM1.2-LR model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.793

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (3rd ed., Vol. 100). Academic Press.

Xin, X., Wu, T., Shi, X., Zhang, F., Li, J., Chu, M., et al. (2019). BCC BCC-CSM2MR model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.1732

Xin, X., Zhang, J., Zhang, F., Wu, T., Shi, X., Li, J., et al. (2018). BCC BCC-CSM2MR model output prepared for CMIP6 CMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.1725

Xu, Y. (2019). Estimates of changes in surface wind and temperature extremes in southwestern Norway using dynamical downscaling method under future climate. *Weather and Climate Extremes*, *26*(1–3), 100234. https://doi.org/10.1016/j.wace.2019.100234

Yukimoto, S., Koshiro, T., Kawai, H., Oshima, N., Yoshida, K., Urakawa, S., et al. (2019a). MRI MRI-ESM2.0 model output prepared for CMIP6 CMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.621

Yukimoto, S., Koshiro, T., Kawai, H., Oshima, N., Yoshida, K., Urakawa, S., et al. (2019b). MRI MRI-ESM2.0 model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.638

Zhang, L., Dobslaw, H., & Thomas, M. (2016). Globally gridded terrestrial water storage variations from GRACE satellite gravimetry for hydrometeorological applications. *Geophysical Journal International*, *206*(1), 368–378. https://doi.org/10.1093/gji/ggw153

Ziehn, T., Chamberlain, M., Lenton, A., Law, R., Bodman, R., Dix, M., et al. (2019a). CSIRO ACCESS-ESM1.5 model output prepared for CMIP6 CMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.2288

Ziehn, T., Chamberlain, M., Lenton, A., Law, R., Bodman, R., Dix, M., et al. (2019b). CSIRO ACCESS-ESM1.5 model output prepared for CMIP6 ScenarioMIP [Dataset]. *Earth System Grid Federation*. https://doi.org/10.22033/ESGF/CMIP6.2291