Reply to Reviewers' comments (Reviewer#2)

Legend
Reviewers' comments
Authors' responses
Direct quotes from the revised manuscript

We thank the reviewer for his time in reading our manuscript and very detailed comments on our manuscript. Point-by-point replies to the comments or suggestions made can be found below. Overall, we have made the following major changes to the manuscript:

- Performed additional analysis using P-E-R and analyzed and compared the results with TWSA-DSI.
- Instead of showing only the ensemble mean of various model and observation-based results, we have now shown the results from individual datasets during the historical period (1985-2014).
- Added extensive discussion about the various mechanisms and governing processes for the observed patterns and the similarities and disparities from the previous studies.

Reviewer #2: The manuscript examines the dry gets drier and wet gets wetter (DDWW) paradigm from a water storage perspective over the terrestrial fraction of the water cycle. The topic is contemplated within the scope of Hydrology and Earth System Sciences via the study of the spatial and temporal characteristics of global water resources and would be of interest to its readership. It is a topic that remains under debate in the last decade and studies that bring new evidence, such as this, can substantially impact the ongoing research. The analysis based on the terrestrial water storage drought severity index is reasonable and aims to quantify wet/dry regime trends under different warming scenarios and assess the agreement or rebuttal of the DDWW paradigm. The study has clear research hypothesis and objectives, that are reflected in the results, and help us increase our understanding of the changes in the global water cycle. However, there are some important aspects that should be addressed before being considered for publication.

Response: We thank the reviewer for his highly encouraging feedback on the manuscript. All the concerns raised have been addressed in the revised manuscript. We hope the modified text, along with the supplementary analyses and discussions, will put forward the results in a much more robust way.

Before moving to the specific revision comments, it is important to note the lack of consistency across the hydrologic and climatic communities in the use of the terms "wet/wetter" and "dry/drier", which in some cases can be misleading (Roth et al. 2021). Thus, it is easy to misinterpret whether a variable is truly appropriate for describing wetting or drying over a region. In this study, a variable (terrestrial water storage) which is not directly involved in the formulation of DDWW paradigm (precipitation/evaporation) is used to validate the paradigm itself. This raises concerns about the applicability of terrestrial water storage as a metric that can confirm or falsify the DDWW hypothesis. The study needs to convincingly prove this. A feasible way to achieve it would be to compare the current findings with P – E, taken from the models (GHMs, LHMs, and CMIP6). This also holds the opportunity to highlight the mechanisms involved in the observed changes and/or pinpoint the biases in the models. Some caution should be taken here since the comparison should also consider surface runoff to satisfy the budget closure. In any case, though, this can help to bridge the different methodologies and explore their complementarity.

Response: We thank the reviewer for cautioning about the interpretation of the topical issue of DDWW and for guiding us through the additional assessment based on the water balance-based metric (i.e., P-E-R). Please find below the parsed-out details about the changes made during the revision of the manuscript.

Inferences of the “wet/wetter” and “dry/drier” terms: We fully agree that the various environmental disciplines define ‘wetter’ and ‘drier’ differently due to the different temporal scales (e.g., from <1 year to >20 years) and spatial extents (i.e., from local to a global extent) within the multidisciplinary climate change communities (Roth et al. 2021). However, the general concept of the terms “wetter”/”dries” imply the increase/decrease of the available amount of freshwater over a specific region. In this case, the TWSA (and TWSA-derived index), constituting the total water mass in the land system, is theoretically reasonable to represent the “wetter” and “drier” global land (Yi et al., 2016; Long et al., 2018). Moreover, we have clarified the meaning, spatial and temporal scales, and the original state that the “wetter” and “drier” were compared to in the methods section as follows.:

TWSA, consisting of the water volume stored in the land surface and subsurface, is applied to define the “wetting” and “drying” conditions of the landmass in this study. The non-dimensional TWS drought severity index (TWS-DSI) is established at both 1°×1° grid cell and regional/global scales, which is normalised by the regional hydroclimatological variability because a given magnitude of TWS deficit could indicate different dryness/wetness conditions in different climate regions. TWS-DSI has clear classification categories based on U.S. Drought Monitor (USDM) and is suitable for comparing dryness/wetness status for different locations and periods (Table S2). It has been widely used in hydrology and climate fields due to its simple structure and effective ability to capture drying and wetting conditions (Pokhrel et al., 2021).

Need for and applicability of TWSA as a metric for examining DDWW: Although the DDWW paradigm was initially formulated using the metric “P-ET” (Held and Soden, 2006; Greve et al., 2014), it is worth noting that independent examinations from multiple aspects, for example, soil moisture (Feng and Zhang, 2015) and runoff (Yang et al., 2019), are attracting increasing attention in the last decade. Therefore, the evaluation of the DDWW paradigm from the TWSA perspective can potentially provide new evidence for the community working on, e.g., ecosystem functioning (Humphrey et al., 2018), sea-level rise (Jeon et al., 2018), water budget (Sheffield et al., 2009), and freshwater availability (Rodell et al., 2018). Moreover, TWSA has not been widely applied as a "wetter" and/or "drier" metric due to the lack of observations globally till the launch of GRACE in 2002. To this end, we have revised the title of the manuscript to prevent using the term "re-examine".

Additional analysis and comparison with 'P-E-R' metric: As suggested by the reviewer, we have additionally established the land water balance metric P-E-R for comparisons with TWS-DSI to highlight the difference between these metrics and to discuss the mechanisms involved. All three constituent variables, i.e., P, ET and R, were taken from the same models (i.e., GHMs, LSMs, and GCMs) as those for TWS-DSI to account for the uncertainty associated with the models and meteorological forcing data. We also used an observation-based combination of P-E-R using precipitation from CRU TS, evapotranspiration from GLEAM, and runoff from GRUN datasets. The inter-comparison between TWS-DSI and P-E-R can help us to bridge the different methodologies and reveal the mechanisms and bias in the changes in dryness/wetness.

References:

Feng, H., Zhang, M., 2015. Global land moisture trends: drier in dry and wetter in wet over land. Sci. Rep. 5, 18018. https://doi.org/10.1038/srep18018

Greve, P., Orlowsky, B., Mueller, B., Sheffield, J., Reichstein, M., Seneviratne, S.I., 2014. Global assessment of trends in wetting and drying over land. Nat. Geosci. 7, 716–721. https://doi.org/10.1038/NGEO2247

Humphrey, V., Zscheischler, J., Ciais, P. et al. Sensitivity of atmospheric CO2 growth rate to observed changes in terrestrial water storage. Nature 560, 628–631 (2018). https://doi.org/10.1038/s41586-018-0424-4

Jeon, T., Seo, K-W., Youm, K., Chen, J., Wilson, C.R. 2018. Global sea level change signatures observed by GRACE satellite gravimetry. Scientific Reports. 8(1): 13519. http://dx.doi.org/10.1038/s41598-018-31972-8.

Long, B., B. Q. Zhang, C. S. He, R. Shao, and W. Tian, 2018: Is there a change from a warm-dry to a warm-wet climate in the Inland River area of China? Interpretation and analysis through surface water balance J. Geophys. Res.: Atmos., 123, 7114–7131, https://doi.org/10.1029/2018jd028436.

Roth, N., Jaramillo, F., Wang-Erlandsson, L., Zamora, D., Palomino-Ángel, S., & Cousins, S. A. (2021). A call for consistency with the terms 'wetter'and 'drier' in climate change studies. Environmental Evidence, 10(1), 1-7.

Sheffield, J., Ferguson, C. R., Troy, T. J., Wood, E. F. and McCabe, M. F. 2009. Closing the terrestrial water budget from satellite remote sensing Geophys. Res. Lett. 36. L07403.

Yang, T., Ding, J., Liu, D., Wang, X., Wang, T., 2019. Combined Use of Multiple Drought Indices for Global Assessment of Dry Gets Drier and Wet Gets Wetter Paradigm. J. Clim. 32, 737–748. https://doi.org/10.1175/JCLI-D-18-0261.1

Yi, S.; Sun, W.K.; Feng, W.; Chen, J.L. Anthropogenic and climate-deiven water depletion in Asia. Geophys. Res. Lett. 2016, 43, 9061–9069.

Rodell, M., Famiglietti, J.S., Wiese, D.N. et al. Emerging trends in global freshwater availability. Nature 557, 651–659 (2018). https://doi.org/10.1038/s41586-018-0123-1

Specific Comments

(1) Lines 1-2: The term "re-examination" should be reconsidered and perhaps be replaced with something like "alternative/complementary examination".

Response: Thank you for the suggestion. After carefully considering the comments from Reviewers #1 and #2, we consider it best to drop the word ‘re-examination’ from the title and update the title to - ‘Global evaluation of the dry gets drier and wet gets wetter paradigm from terrestrial water storage changes perspective’.

(2) Lines 25-27: It would be more appropriate to reference the original study of Held and Soden (2006).

Response: Thank you for the suggestion. We have referenced the original study.

(3) Lines 34-42: You might want to look at the work of Roderick et al. (2014).

Response: We have read through the suggested key study and have included relevant information in our manuscript.

(4) Lines 43-74: This paragraph could be split into two (line 56 perhaps?) to improve readability.

Response: We have split the paragraph as suggested.

(5) Line 84 (Table 1): Since the three GRACE reconstructions come from two papers it could help the readers if they were also somehow distinguished in the table.

Response: Only the GRACE (CSR) reconstruction from Li et al. (2021) is used for the evaluation of the DDWW paradigm in the new version owing to high reliability and robustness (produced using a combination of three data-driven approaches and compared against various hydroclimatic indices and water storage component outputs by global hydrological models) of this product compared to the other ones (Li et 2020; 2021). We have modified the table accordingly.

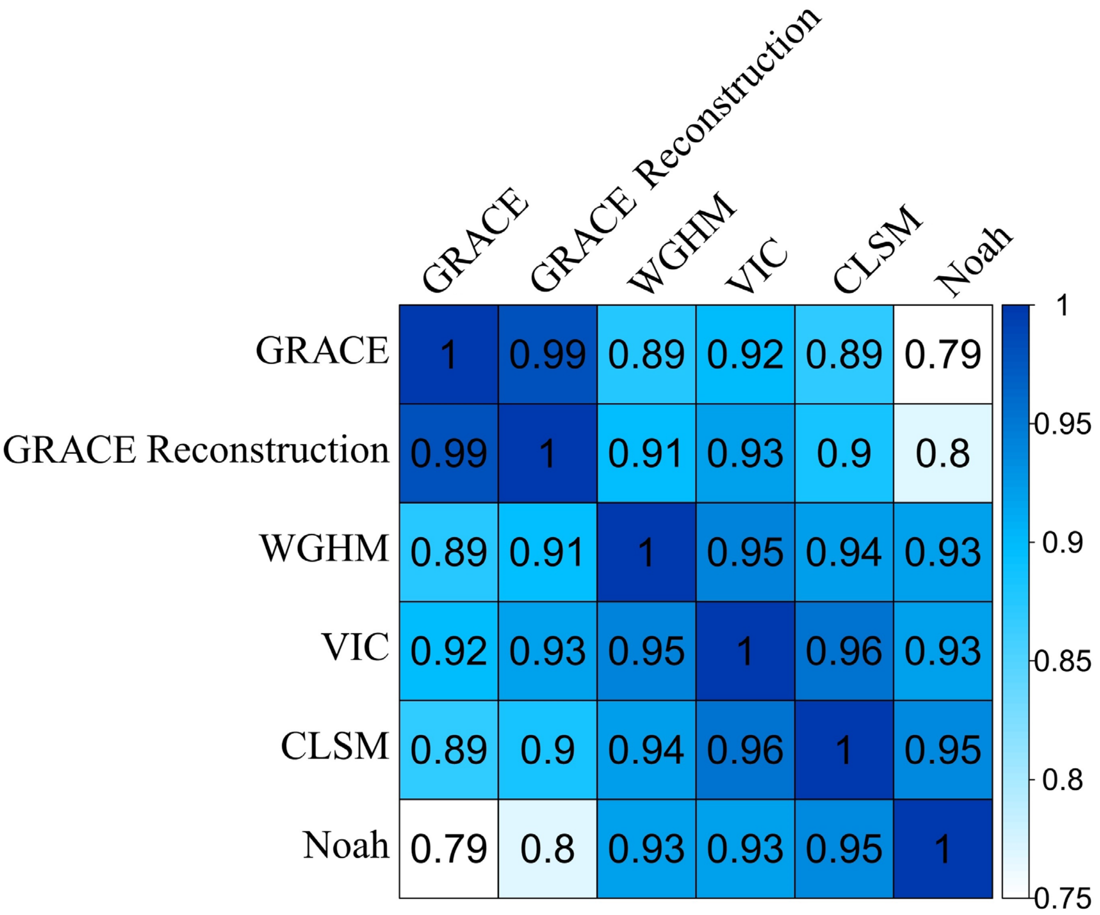
References:

Li, F., Kusche, J., Chao, N., Wang, Z., Loecher, A., 2021. Long-Term (1979-Present) Total Water Storage Anomalies Over the Global Land Derived by Reconstructing GRACE Data. Geophys. Res. Lett. 48, e2021GL093492. https://doi.org/10.1029/2021GL093492

Li, F., Kusche, J., Rietbroek, R., Wang, Z., Forootan, E., Schulze, K., Lück, C. 2020. Comparison of Data-driven Techniques to Reconstruct (1992-2002) and Predict (2017-2018) GRACE-like Gridded Total Water Storage Changes using Climate Inputs[J]. Water Resources Research, 56(5), e2019WR026551. https://doi.org/10.1029/2019wr026551.

(6) Lines 84-85: Averaging the datasets always comes with certain challenges. For example, in this case we would expect that the three reconstructions of GRACE are strongly correlated and thus their impact to the estimation of the mean would be stronger. A cross-correlation matrix between the datasets would help to assess the magnitude of the impact and comment on it in the manuscript.

Response: Since different models/products have different TWSA definitions, we opt to demonstrate the individual examination results of different datasets rather than merely the ensemble mean. The additional cross-correlation matrix with the GRACE observations shows the reasonable consistency of the datasets with the Pearson correlation coefficient ranging from 0.79 (Noah) to 0.99 (GRACE reconstruction). We have added this information in the revised manuscript.



	GRACE	GRACE Reconstruction	WGHM	VIC	CLSM	Noah
GRACE	1	0.99	0.89	0.92	0.89	0.79
GRACE Reconstruction	0.99	1	0.91	0.93	0.9	0.8
WGHM	0.89	0.91	1	0.95	0.94	0.93
VIC	0.92	0.93	0.95	1	0.96	0.93
CLSM	0.89	0.9	0.94	0.96	1	0.95
Noah	0.79	0.8	0.93	0.93	0.95	1

Figure R1. Correlation matrix between the DATASET members and GRACE over global land excluding Antarctica and Greenland during April 2002-December 2014. Note: The numbers mean the Pearson correlation coefficients within two variables.

(7) Lines 87-88: Can you please provide some information about the percentage of the missing values, as well as their distribution among the number of consecutive missing values?

Response: We have provided information about the percentage and distribution of the missing months in the revised manuscript, as below.

The missing months (12% of the months, i.e., June 2002, July 2002, June 2003, January 2011, June 2011, May 2012, October 2012, March 2013, August 2013, September 2013, February 2014, July 2014, December 2014) of GRACE measurements have been filled using a linear interpolation method.

(8) Lines 107-108: “kinds” might not be necessary here.

Response: We have removed “kinds” from both the places as suggested.

(9) Lines 167-168: Please also cite the original paper of Hempel et al. (2013) which has been used by Xiong et al. (2022).

Response: We have also cited the primary reference as suggested.

(10) Lines 177-178: Before applying linear interpolation and assessing the statistical significance of the slope the auto-correlation structure of the time series should be investigated. High values of auto-correlation coefficient could result to biased estimates of t, so if this is the case, alternative methods of slope significance should be applied (Hamed and Rao, 1998; Yue et al. 2002). In addition, the reference to the work of Greve et al. (2014) should be revisited as a different statistical test is applied at that study than the t-test.

Response: We have investigated the first-order autocorrelation structure of the time series of multiple datasets and the CMIP6 GCMs using the Durbin-Watson test (Figure R2) (Durbin and Watson, 1950, 1951). A total of 20% (GRACE reconstruction), 43% (WGHM), 41% (VIC), 23% (CLSM), 29% (Noah), and 20% (GCM) of the grid cells do not present autocorrelation during the historical period 1985-2014. For the future period, the percentage is 25%, 26%, and 22% under the SSP126, SSP245, and SSP585 scenarios, respectively. In this case, we select the modified Mann-Kendall trend test to obtain the true variance under the autocorrelation structure displayed (Hamed and Rao, 1998). In addition, we have changed the reference to the work of Greve et al. (2014) in the revised manuscript.

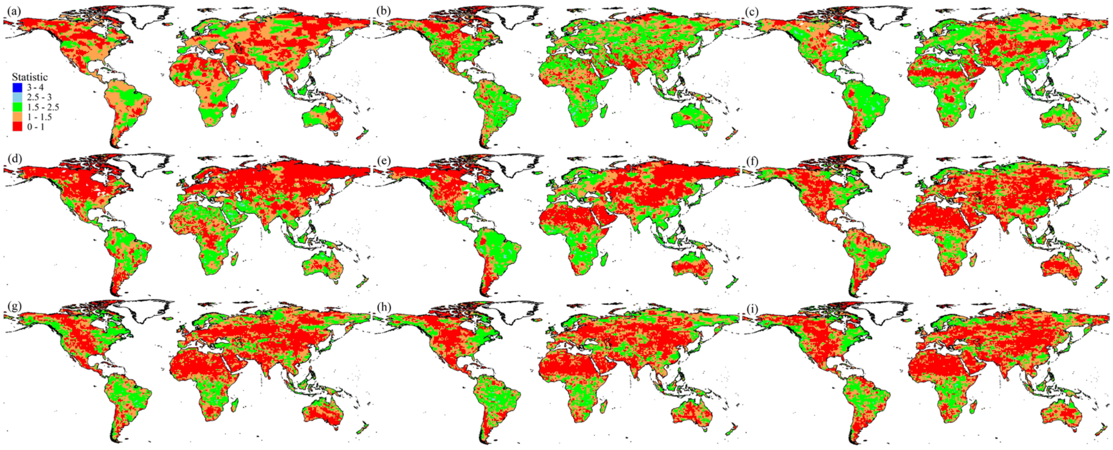


Figure R2. Global assessment of the autocorrelation during the (a-f) historical (1985-2014) and future (2071-2100) period under (b) SSP126, (c) SSP245, and (d) SSP585 scenarios. Note: The historical results are based on the (a) GRACE reconstruction, (b) WGHM, (c) VIC, (d) CLSM, (e) Noah, and (f) ensemble mean of eight GCMs, respectively using the Durbin-Watson test. The future results are based on the ensemble of eight GCMs. Generally, the residuals are considered not correlated when the Durbin-Watson test statistic has a value between 1.5 and 2.5. If the statistic is below 1 or above 3, then there is definitely autocorrelation among the residuals.

Reference:

Durbin, J., Watson, G. S. 1950. Testing for Serial Correlation in Least Squares Regression, I. Biometrika 37 : 409-428.

Durbin, J., Watson, G. S. 1951. Testing for Serial Correlation in Least Squares Regression, II. Biometrika 38: 159-179.

Hamed, K. H., Rao, A. R. 1998. A modified Mann-Kendall trend test for autocorrelated data. Journal of hydrology, 204(1-4), 182-196.

(11) Line 180: Should the region be considered as "uncertain" or should it be considered a region with no or non-significant long-term change?

Response: We have changed the terminology from "uncertain" to the "non-significant" regions throughout the revised manuscript.

(12) Lines 183-184: It is preferable to keep a single tense for the whole manuscript (past/present). This comment also applies for other lines.

Response: We regret the non-uniformity. We have kept the present tense for the whole of the revised manuscript.

(13) Line 185: Citation of the report is missing.

Response: Since we have excluded the case analysis based on the IPCC AR6 SREX regions in the revised manuscript according to the suggestions from Reviewer #1, the citation of the report is not needed anymore.

(14) Lines 190-208: My understanding after reading the bias-correction method of Xiong et al. (2022) is that the CMIP6 ensemble was bias-corrected using GRACE TWSA. If this holds true, then the evaluation of the TWSA derived from the ensemble mean of CMIP6 raises some questions about its validity and therefore NRMSE is lower compared to DATASET.

Response: In addition to the comparison to the GRACE TWSA, we also provided the evaluation of the bias-corrected TWSA changes against the water balance estimates (i.e., P-E-R) during 1985-2014 (Figures R3 and R4). The observation-based water balance estimates correlate well with GRACE TWSA and GCM-modelled P-E-R with a correlation coefficient of 0.62 and 0.93, respectively. The GCM-simulated changes in TWSA also present a strong correlation with the observational products before and after bias correction. The spatial distribution of correlation coefficients between TWSC from observations and GCMs with and without bias correction shows the performances in regions with good accuracy, like Alaska, western parts of the Tibetan Plateau, and northern Russia, decrease after bias correction, which might be caused by the simplified treatment of permafrost in GCMs due to the prevailing uncertainties in, e.g., changes in thermophysical properties of the soil during freezing and thawing cycles (Burke et al., 2020). On the contrary, the areas with relatively poorer accuracy before bias correction, such as North Africa and northern South America, slightly improve after bias correction. Notwithstanding the observed differences in some regions, our trend-preserving method used for bias correction would not influence the long-term trend estimations of both TWSA and TWS-DSI and therefore does not impact our evaluation of the DDWW paradigm (Hempel et al., 2013).

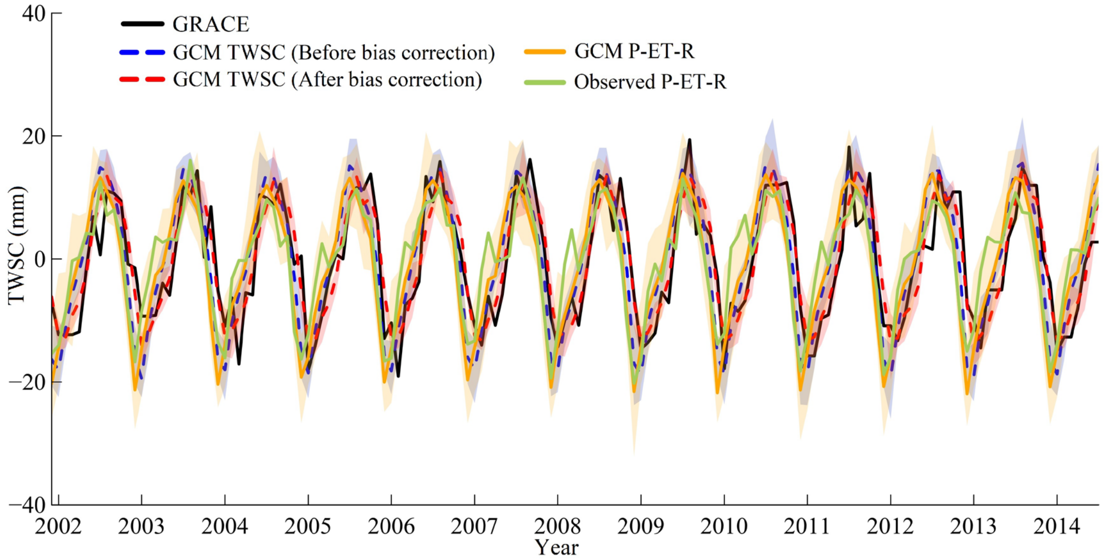


Figure R3. Time series of the monthly changes in TWSA (TWSC) and water balance estimates (i.e., P-E-R) derived from GRACE, GCM, and observations during 2002-2014. Note: The shaded regions represent the spread of the CMIP6 ensemble.

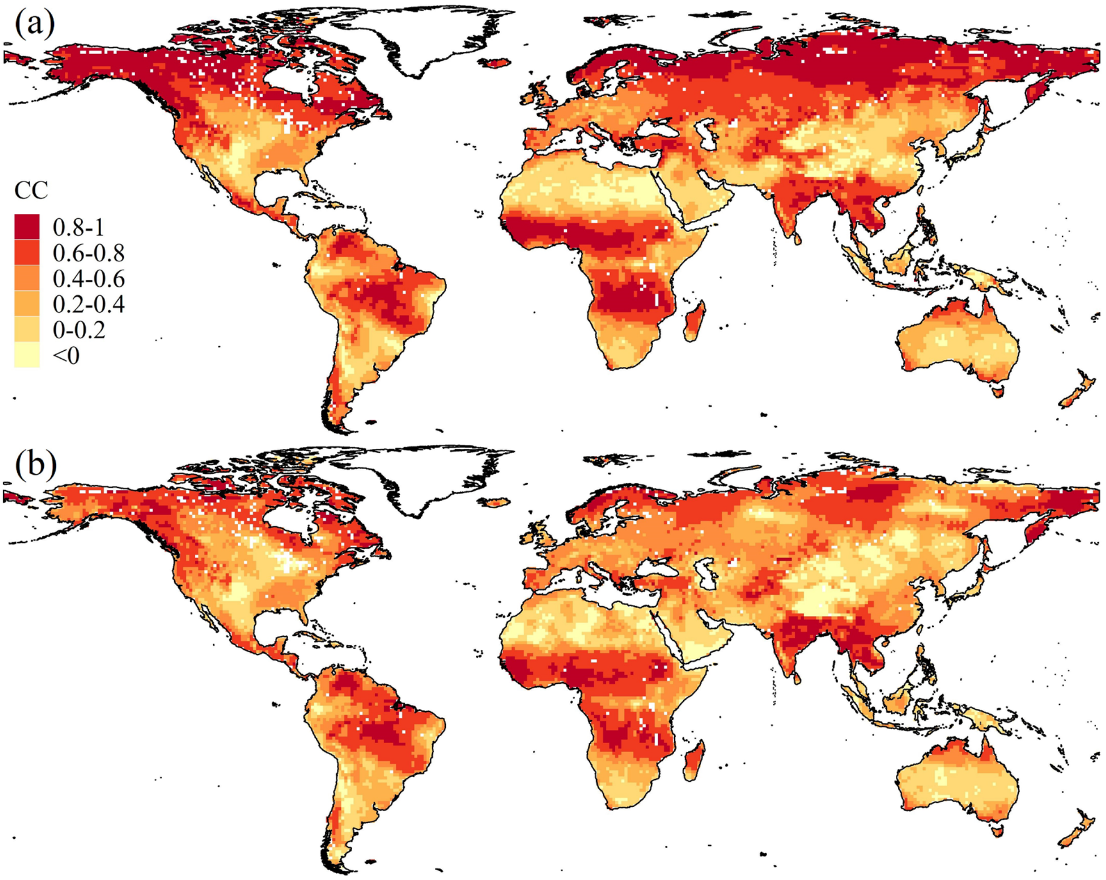


Figure R4. Spatial distribution of correlation coefficient between monthly water balance estimates of TWSA changes and the ensemble mean of GCM data (a) before and (b) after bias corrections during 1985-2014. The blank grids indicate the missing values of the datasets.

References:

Burke, E.J., Zhang, Y., Krinner, G. 2020. Evaluating permafrost physics in the coupled model intercomparison project 6 (CMIP6) models and their sensitivity to climate change. Cryosphere., 14 (9) , pp. 3155-3174

Hempel, S., Frieler, K., Warszawski, L., Schewe, J., Piontek, F. 2013. A trend preserving bias correction–the ISI-MIP approach. Earth System Dynamics, 4(2), 219-236.

(15) Lines 214-216: Is there any likely explanation about the increase in the range of DATASET after 2010? Perhaps it can be linked to the decline of TWSA of a specific dataset.

Response: The reviewer is correct in his anticipation. This is caused by the abrupt changes in TWSA from the PCR-GLOBWB model. In the revised manuscript, since 1) we have excluded the usage of this model (because we attempt to construct the metric P-E-R for parallel comparison and it is not available for the PCR-GLOBWB model) and 2) opt to present the individual members of different datasets instead of just the ensemble mean, the increase in the range of DATASET after 2010 has disappeared (Figure R5).

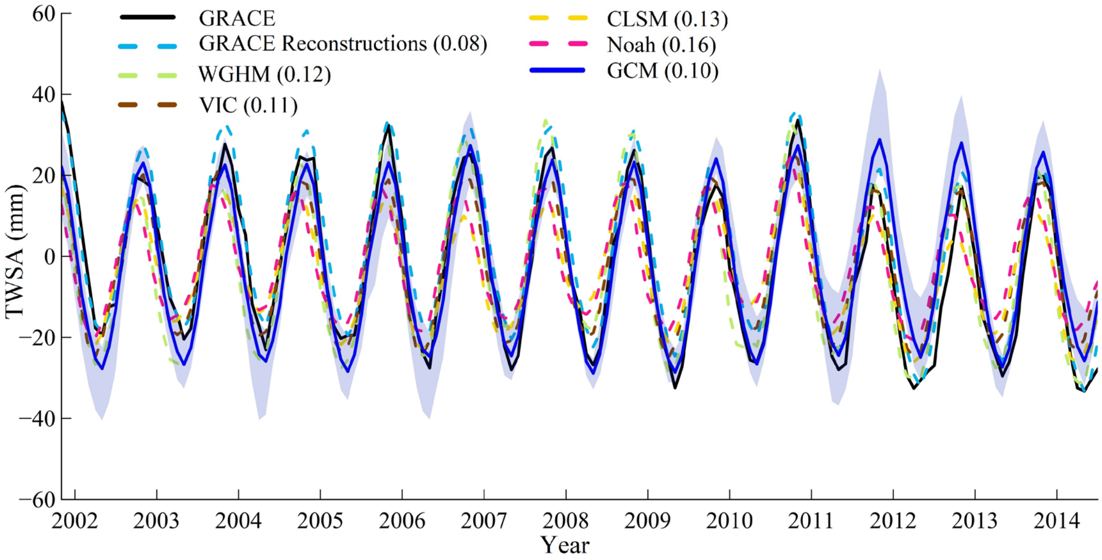


Figure R5. Time series of monthly TWSA derived from the GRACE products and different TWSA datasets over the global land excluding Antarctica and Greenland during the period April 2002-December 2014. Note: NRMSE between GRACE and different datasets are also shown. The deep blue line denotes the ensemble mean of eight GCMs. The shaded areas represent the range of TWSA values among the individual GCM datasets.

(16) Line 219: It would be helpful to remind the readers that for the historical period DATASET is used, and not only the CMIP6 data that are used for the scenarios.

Response: We have clarified the data source for the trend estimate for the historical and future periods in both text and figure captions again.

(17) Line 221: Since you are referring to the slope "/a" is redundant. Still if you would like to keep it, you could consider replacing it with "/yr".

Response: We have replaced it with “/yr” throughout the revised manuscript.

(18) Line 228: I think “increasing” is the correct word here.

Response: Thank you. We have updated as suggested.

(19) Lines 228-229: There is no study related to S. Europe in the references. Most importantly, in Figure 1b it appears that SE. Europe is wet and the rest of the south not significantly drier. This contradicts older studies reporting a drying trend over the Mediterranean (e.g., Hoerling et al. 2012) and could shed new light to the ongoing discussion about the current and future conditions of S. Europe, so I would recommend elaborating more.

Response: As indicated by the reviewer, the reduction in winter precipitation has become a regular phenomenon in the Mediterranean and caused increased seasonal drought risks (Hoerling et al. 2012, Wagner et al., 2019). However, no statistically significant long-term decreasing trends are detected at the annual timescale for the region (Peña-Angulo et al., 2020; Vicente-Serrano et al., 2020). In particular, the centre of the Mediterranean basin including France, Italy, Croatia, and Slovenia presented insignificant decreasing trends in annual precipitation, while the western and eastern regions such as Spain and Greece demonstrated increasing trends during 1991-2018 (Caloiero et al., 2018; Peña-Angulo et al., 2020). These findings are generally consistent with our results based on TWS-DSI derived from different models during the historical period 1985-2014 (Figures R6 and R7).

We have revised this statement and added references related to the drying/wetting trends over South Europe. Moreover, we have added discussions for the past and future changes in wetness/dryness over southwestern Europe in the revised version.

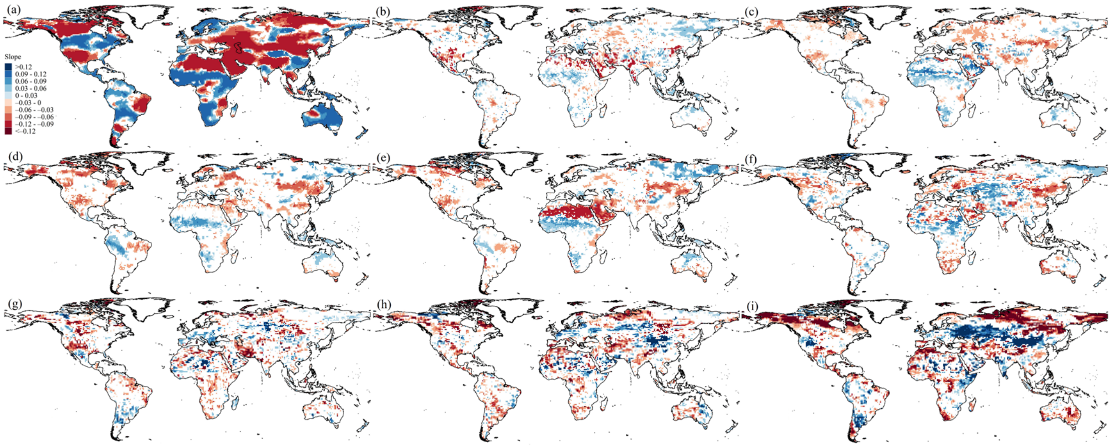


Figure R6. Global distribution of the significant (p<0.05) long-term trends in TWS-DSI during (a-f) the historical (1985-2014) and future (2071-2100) period under (g) SSP126, (h) SSP245, and (i) SSP585 scenarios. Note: The historical results are based on the (a) GRACE reconstruction, (b) WGHM, (c) VIC, (d) CLSM, (e) Noah, and (f) ensemble mean of eight GCMs, respectively. The future results are based on the ensemble of eight GCMs.

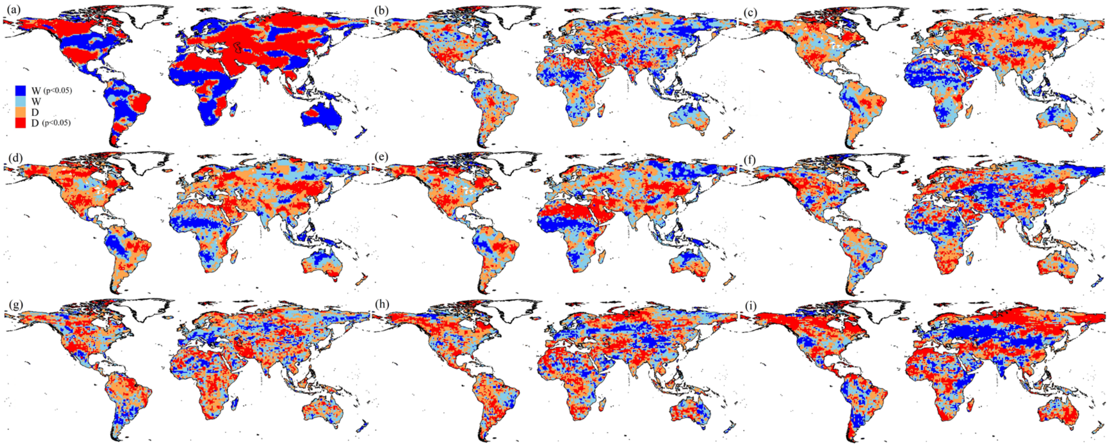


Figure R7. Global distribution of the classification in long-term trends in TWS-DSI during (a-f) the historical (1985-2014) and future (2071-2100) period under (g) SSP126, (h) SSP245, and (i) SSP585 scenarios. Note: The historical results are based on the (a) GRACE reconstruction, (b) WGHM, (c) VIC, (d) CLSM, (e) Noah, and (f) ensemble mean of eight GCMs, respectively. The future results are based on the ensemble of eight GCMs. "D" and "W" indicate regions with drying and wetting trends, respectively.

Reference:

Caloiero T, Veltri S, Caloiero P, Frustaci F. 2018. Drought analysis in Europe and in the Mediterranean Basin using the standardized precipitation index. Water 10(8):1043

Hoerling, M., Eischeid, J., Perlwitz, J., Quan, X., Zhang, T., Pegion, P. 2012. On the increased frequency of Mediterranean drought. Journal of climate, 25(6), 2146-2161.

Peña-Angulo, D., Vicente-Serrano, S. M., Domínguez-Castro, F., Murphy, C., Reig, F., Tramblay, Y., Trigo, R. M., Luna, M. Y., Turco, M., Noguera, I., Aznárez-Balta, M., García-Herrera, R., Tomas-Burguera, M., and El Kenawy, A.: Long-term precipitation in Southwestern Europe reveals no clear trend attributable to anthropogenic forcing, Environ. Res. Lett., 2020. 15, 094070, https://doi.org/10.1088/1748-9326/ab9c4f.

Vicente-Serrano, S.M., Domínguez-Castro, F., Murphy, C., Hannaford, J., Reig, F., Peña-Angulo, D., Tramblay, Y., Trigo, R.M., Mac Donald, N., Luna, M.Y., Mc Carthy, M., Van der Schrier, G., Turco, M., Camuffo, D., Noguera, I., García-Herrera, R., Becherini, F., Della Valle, A., Tomas-Burguera, M., El Kenawy, A. 2021. Long-term variability and trends in meteorological droughts in Western Europe (1851–2018). Int. J. Climatol., 41, pp. E690-E717

Wagner, B., Vogel, H., Francke, A., Friedrich, T., Donders, T., Lacey, J. H., et al. 2019. Mediterranean winter rainfall in phase with African monsoons during the past 1.36 million years. Nature, 573(7773), 256–260. https://doi.org/10.1038/s41586-019-1529-0

(20) Lines 233-235: Do you mean that your results for these regions disagree with the previous studies? If yes, you could clarify a bit and discuss potential reasons for the disagreement. Also, you might want to replace "alternatively" with "on the contrary".

Response: We regret the misinterpretation. Our results are consistent with the previous studies and we have revised this statement from "wet to dry" to "dry to wet" in the new version. We also replaced "alternatively" with "on the contrary". The revised sentence is as follows:

On the contrary, some regions, such as the Amazon River basin, south Africa and eastern Australia, presenting wetting trends, are considered to experience a climatic shift from dry to the wet period (Chen et al., 2010; Gaughan and Waylen, 2012).

(21) Line 236: In SSP126 scenario, S. Europe also has a strong wetting trend.

Response: We have corrected this statement and elaborated more about the future

conditions over southwestern Europe in the revised version as follows:
Specifically, all three scenarios confirm the significant (p<0.05) wetting trends in North China, South Mongolia, central Asia, northern border of Canada, and South Europe, with the increase in the intensity and spread along with the enhancement of climate scenarios (Figures 1, 2, S14, and S15).

(22) Lines 236-245: This paragraph discusses only the SSP126 scenario, while the other two more probable scenarios remain uncommented. It would be nice to discuss the differences between each projection scenario and highlight the regions that all scenarios agree. Another striking difference appears in spatial clustering between the historical period and the model results in terms. It is evident that in the historical period there is stronger spatial homogeneity, while the models replicate this behavior only for SSP126 scenario. Any idea why this is happening?

Response: Thank you for the informative suggestion. We have added descriptions for the future projections under various scenarios and their similarities and differences in the revised version as follows:
Specifically, all three scenarios confirm the significant (p<0.05) wetting trends in North China, South Mongolia, central Asia, northern border of Canada, and South Europe, with the increase in the intensity and spread along with the enhancement of climate scenarios (Figures 1, 2, S14, and S15). Similarities are found in the drying trends in the majority of Russia, northern North America, and South Africa. The wetting trends are apparently caused by the increase in precipitation (Figure S16) (Milly et al., 2005; Seneviratne et al., 2006). The arid Arab region is also projected to become wetter because of the increase in precipitation and the decrease in evapotranspiration. On the contrary, the drying trends are mainly controlled by the rapidly intensifying evapotranspiration in a warming climate (Figure S17) (Allen et al., 2010; Vicente-Serrano et al., 2010), with the precipitation and runoff slightly increasing (Figures S16 and S18). The obvious drying trend around Canada's subarctic lakes is attributed to the high vulnerability to droughts when snow cover declines under increasing temperature (Bouchard et al., 2013). However, there exist scenario-variable divergences over the continents of South America, Australia, India, and the Mediterranean basin, which are generally caused by the various patterns in precipitation under different scenarios with the increasing evapotranspiration over there. The runoff also follows the patterns of precipitation but with comparably lesser magnitudes. The above-cited figures (Figures 1, 2, and S14-S18) are provided below for better comprehension.
Since the previous ensemble mean results of DATASET are mainly affected by three GRACE reconstructions (they are highly correlated with each other), the spatial distribution of trends shows spatial clustering. By presenting the individual results of each subset (Figure 1 below), we can only see this pattern in GRACE reconstruction from UTCSR, whereas the GHMs, LSMs, and GCMs show different spatial characteristics. We also mention this in the revised manuscript. Furthermore, we also show the differences between the DDWW results in previous and present manuscripts in Table R1, suggesting the main conclusions are unaffected.

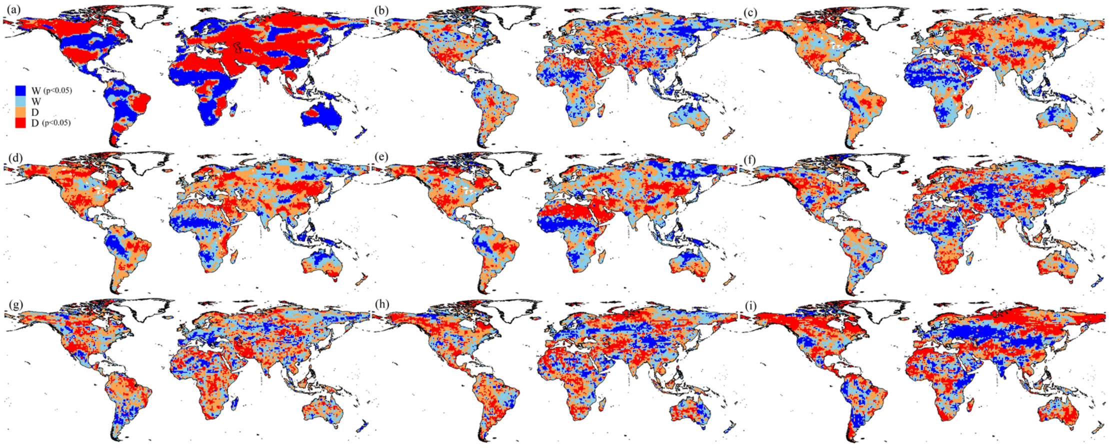


Figure 1. Global distribution of the classification in long-term trends in TWS-DSI during (a-f) the historical (1985-2014) and future (2071-2100) period under (g) SSP126, (h) SSP245, and (i) SSP585 scenarios. Note: The historical results are based on the (a) GRACE reconstruction, (b) WGHM, (c) VIC, (d) CLSM, (e) Noah, and (f) ensemble mean of eight GCMs, respectively. The future results are based on the ensemble of eight GCMs. "D" and "W" indicate regions with drying and wetting trends, respectively.

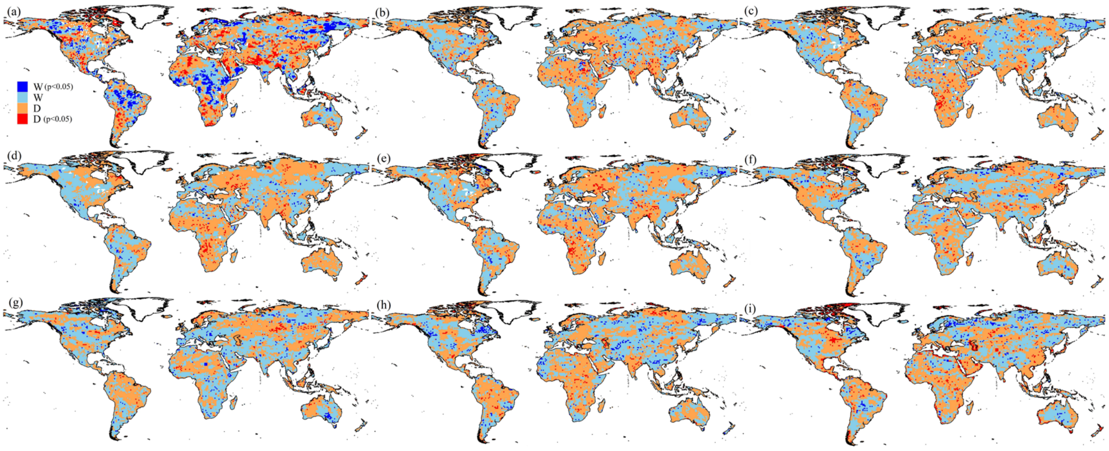


Figure 2. Global distribution of the classification in long-term trends in P-E-R during (a-f) the historical (1985-2014) and future (2071-2100) period under (g) SSP126, (h) SSP245, and (i) SSP585 scenarios. Note: The historical results are based on the (a) observation-based products (i.e., CRU P, GLEAM E, and GRUN R), (b) WGHM, (c) VIC, (d) CLSM, (e) Noah, and (f) ensemble mean of eight GCMs, respectively. The future results are based on the ensemble of eight GCMs. "D" and "W" indicate regions with drying and wetting trends, respectively.

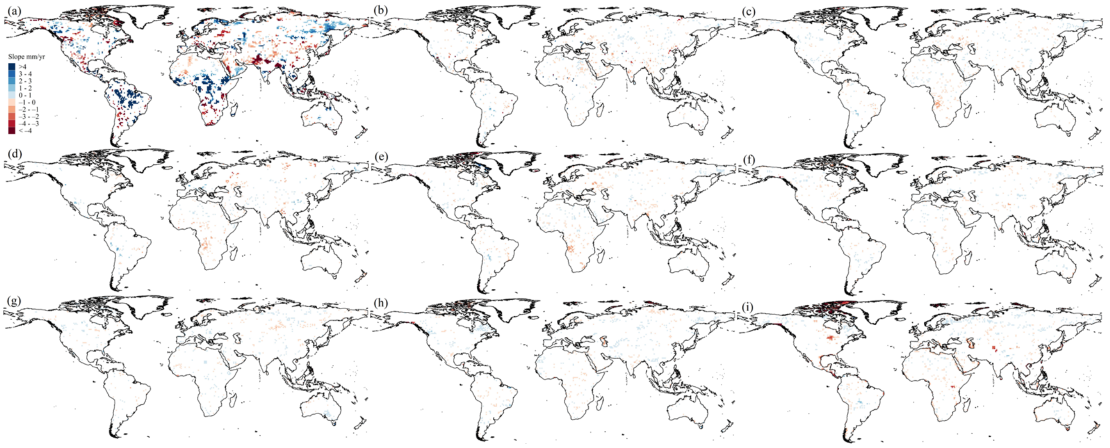


Figure S15. Global distribution of the significant ($p<0.05$) long-term trends in P-E-R during (a-f) the historical (1985-2014) and future (2071-2100) period under (g) SSP126, (h) SSP245, and (i) SSP585 scenarios. Note: The historical results are based on the (a) observational products (i.e., CRU P-GLEAM E-GRUN R), (b) WGHM, (c) VIC, (d) CLSM, (e) Noah, and (f) ensemble mean of eight GCMs, respectively. The future results are based on the ensemble of eight GCMs.

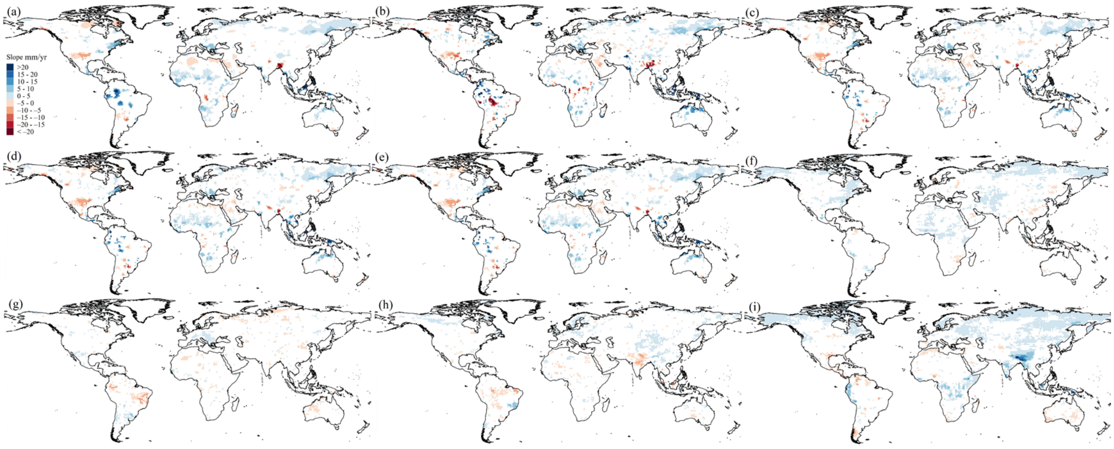


Figure S16. Global distribution of the significant ($p<0.05$) long-term trends in P during (a-f) the historical (1985-2014) and future (2071-2100) period under (g) SSP126, (h) SSP245, and (i) SSP585 scenarios. Note: The historical results are based on the (a) CRU, (b) WGHM, (c) VIC, (d) CLSM, (e) Noah, and (f) ensemble mean of eight GCMs, respectively. The future results are based on the ensemble of eight GCMs. The VIC, CLSM, and Noah models are forced by the same precipitation dataset because they are from the GLDAS 2.0 family.

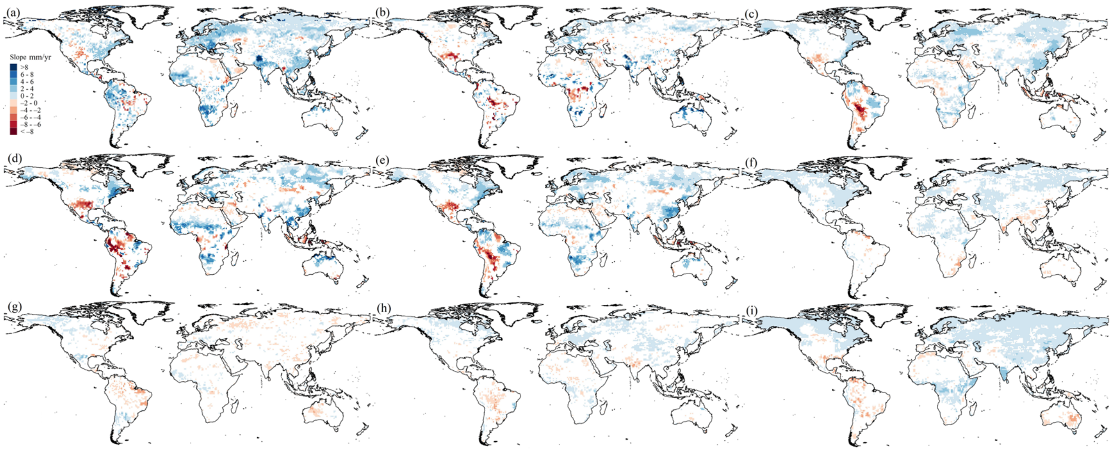


Figure S17. Global distribution of the significant ($p<0.05$) long-term trends in E during (a-f) the historical (1985-2014) and future (2071-2100) period under (g) SSP126, (h) SSP245, and (i) SSP585 scenarios. Note: The historical results are based on the (a) GLEAM E, (b) WGHM, (c) VIC, (d) CLSM, (e) Noah, and (f) ensemble mean of eight GCMs, respectively. The future results are based on the ensemble of eight GCMs.

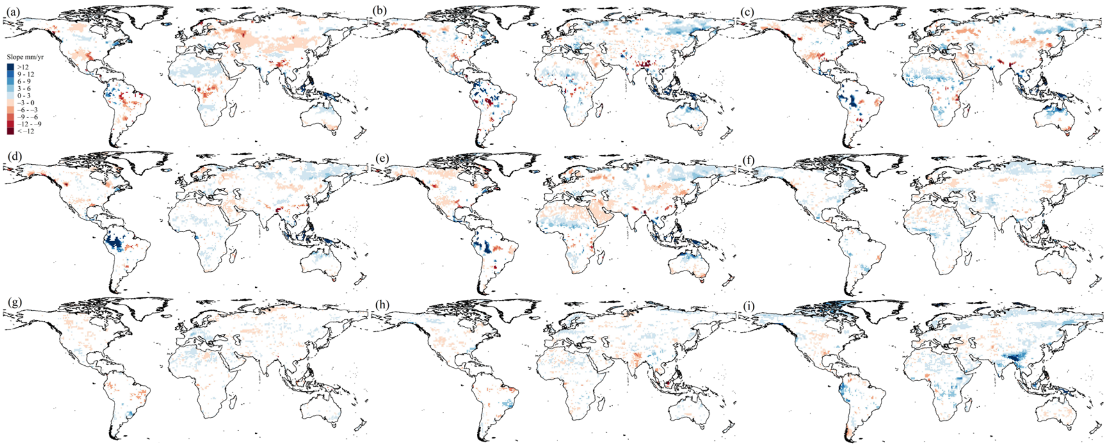


Figure S18. Global distribution of the significant ($p<0.05$) long-term trends in R during (a-f) the historical (1985-2014) and future (2071-2100) period under (g) SSP126, (h) SSP245, and (i) SSP585 scenarios. Note: The historical results are based on the (a) GRUN, (b) WGHM, (c) VIC, (d) CLSM, (e) Noah, and (f) ensemble mean of eight GCMs, respectively. The future results are based on the ensemble of eight GCMs.

Table R1. Summary of the changes in the DDWW test results over global land during 1985-2014.

Model/dataset	Previous results (ensemble mean of DATASET)	Updated results (individual datasets) [range]	Remark
DD	16.7%	6.47%-20.17%	From the perspective of TWSA, the DDWW is still challenged based on both the ensemble mean (previous version) and the individual datasets (current version) used in this study.
DW	8.4%	5.42%-16.13%	
WW	11.4%	4.54%-20.67%	
WD	14.9%	4.79%-19.3%	
TD	2.1%	0.95%3.88%	
TW	1.8%	0.73%-2.63%	
Non-significant	45.1%	17.2%-72.42%	

(23) Line 245: You could consider rephrasing to “a pattern also considered”.

Response: Rephrased as suggested.

(24) Lines 246-259: It is not very clear what the investigation of the changes over the SREX regions offers to the study.

Response: We have removed the case analysis based on the IPCC AR6 SREX regions in the revised manuscript while providing the processed data for the community to use in any regions (e.g., SREX, basins, etc.) of interest.

(25) Line 260: The legend is not very clear (some spaces between the numbers would help). Also, the scale order should be from higher to lower. Since the stippling marks are not very clear, could you please remove them and reproduce this map with only the statistically significant slopes in the supplementary material?

Response: We have revised the figures according to your suggestions in the revised version (Figures R6 and R7 above).

(26) Line 266: Same concerns here for the legend as the previous comment on Figure 1. The bar plot needs also revising as D should be above D ($p<0.05$). Additionally, I am not certain that the pie plots help the readers and are not discussed in the manuscript. You might want to consider removing them and adding them as a separate figure in the supplementary material.

Response: We have removed them for clarity in the revised manuscript as suggested.

(27) Lines 273-276: These lines would fit better to the Methods section.

Response: We have moved this paragraph to the Methods section as suggested.

(28) Line 274: Please elaborate about transitional regions.

Response: We have provided more information about the transitional regions in the revised manuscript as follows:

An approximate 7.9% of the land area is defined as the transitional region, referring to an intermediate between arid and humid climates. The transitional region generally lies in the shared boundaries of the humid and arid regions (e.g., western America, northern Canada, central Asia, western Africa, East Russia, and Australia).

(29) Line 290: You might want to remove “Under climate change” since you mention the SSP126 scenario.

Response: Removed.

(30) Lines 295-304: Again, I have similar concerns about SREX regions as the ones for lines 246-259. If you decide to keep them and justify the added value they offer in the analysis, please consider presenting these results in an individual paragraph.

Response: We have removed this content in the revised manuscript and have restrained our analysis on the global land only with additional regional analysis about the selected region of the Qinghai-Tibetan Plateau. Since we have provided the processed data publicly available, further studies may focus on different regions of interest.

(31) Line 319: “In climate model projections”, would be more appropriate than “Under climate change”.

Response: We have changed it as suggested.

(32) Line 336: Please see the comment about Figure 2 (Line 266) regarding the pie charts.

Response: We have removed this figure according to the suggestions from Reviewer #1.

(33) Line 343: I would recommend using “Non-significant” instead of “Uncertain” here.

Response: We have used the term “Non-significant” throughout the revised manuscript.

(34) Line 348: It would be helpful to the readers to link the limitations with some suggestions for future research, especially for the first two paragraphs.

Response: As suggested, we have linked the inherent limitations of this study with future research (mainly related to the use of advanced bias correction methods, and including a larger number of GCM outputs as and when they are available) in the new version as follows:

…Overall, the models with completed TWS components are more suitable for assessing the TWSA changes at the global scale for future research, such as the continuously developing hyper-resolution global hydrological models (e.g., WGHM), which can help to avoid the uncertainty associated with the lack of key TWSA elements in most LSMs (e.g., surface water and groundwater) (Pokhrel et al., 2021).

…Advanced bias-correction methods (e.g., Lange, 2019 and Francois et al., 2020) might play critical roles in reducing such errors in meteorological variables for future hydrologic impact studies, especially when combined with the start-of-the-art GHMs and LSMs as mentioned above. The inclusion of more GCMs can also help to estimate the uncertainties in the meteorological inputs in climate change scenarios. Although it is challenging to explicitly attribute and quantify these uncertainties in the absence of a ‘true’ reference observation dataset, the ensemble averaging method has been used to integrate the multi-source TWSA data. Moreover, since the meaning and hence the results and interpretation of the 'dry' and 'wet' varies across disciplines, land or ocean, target variable(s), and the problem in question (Roth et al., 2021), future studies may focus on various spatial (e.g., local, regional, basin, zonal averages) and temporal (monthly, seasonal, annual) scales using our processed data with additional model outputs (e.g., more number of GCMs).

(35) Lines 379-385: Similarly to lines 190-208, bias-correction comes with certain limitations which need to be mentioned here.

Response: Since we have conducted independent comparisons between bias-corrected CMIP6 TWSA changes and observation-based water balance estimates (i.e., P-E-R), we think such limitations of the bias correction method using the GRACE data may not exist anymore.

(36) Line 382: A minor typo here “bias correction”.

Response: We regret the typo. We have corrected it as suggested.

(37) Lines 403-407: It would be preferable to present the differences to the 0.05 threshold both in text and Figure S13.

Response: We have provided the difference between the 0.05 significance level to the 0.01 and 0.1 thresholds both in the text and supplementary file (same as Tables R1 and R2 below).

Table R1. Differences between the DDWW test results at 0.01 and 0.05 significance levels.

Model/dataset	GRACE Reconstructions	WGHM	VIC	CLSM	Noah	GCM (historical period)	GCM (SSP126)	GCM (SSP245)	GCM (SSP585)
DD	-0.98%	-2.06%	-2.78%	-3.86%	-3.40%	-2.92%	-2.91%	-3.00%	-2.73%
DW	-0.84%	-2.36%	-3.19%	-2.24%	-2.45%	-2.68%	-2.46%	-2.40%	-2.74%
WW	-2.23%	-3.93%	-2.79%	-3.34%	-3.62%	-2.92%	-2.70%	-2.19%	-2.63%
WD	-1.81%	-2.43%	-4.04%	-3.99%	-3.31%	-3.94%	-3.79%	-4.64%	-4.00%
TD	-0.27%	-0.36%	-0.64%	-0.67%	-0.52%	-0.49%	-0.58%	-0.59%	-0.57%
TW	-0.30%	-0.52%	-0.41%	-0.52%	-0.58%	-0.54%	-0.45%	-0.39%	-0.54%
Non-significant	6.42%	11.65%	13.84%	14.61%	13.89%	13.48%	12.88%	13.21%	13.20%

Table R2. Differences between the DDWW test results at 0.1 and 0.05 significance levels.

Model/dataset	GRACE Reconstruction	WGHM	VIC	CLSM	Noah	GCM (historical period)	GCM (SSP126)	GCM (SSP245)	GCM (SSP585)
DD	0.52%	1.46%	1.97%	2.14%	2.23%	1.82%	1.85%	1.89%	1.60%
DW	0.48%	1.75%	1.85%	1.42%	1.54%	1.79%	1.71%	1.59%	1.54%
WW	1.13%	3.12%	2.19%	2.04%	2.24%	2.16%	2.07%	1.74%	1.72%
WD	0.98%	1.98%	2.64%	2.94%	2.39%	2.80%	2.76%	3.17%	2.45%
TD	0.20%	0.45%	0.44%	0.73%	0.35%	0.43%	0.46%	0.36%	0.43%
TW	0.09%	0.38%	0.37%	0.39%	0.46%	0.30%	0.31%	0.31%	0.31%
Non-significant	-3.39%	-9.14%	-9.47%	-9.66%	-9.22%	-9.28%	-9.16%	-9.06%	-8.04%

(38) Lines 421-424: A quite strong statement appears here. Are there any other studies that support it or is it derived only by the results of this study?

Response: Because we have presented the individual results of each subset of the DATASET and compared them with the corresponding metric P-E-R, we have weakened this statement in the new version as follows:

Given the inherent magnitude bias from various GCMs projections, the ensemble averaging method has the potential to provide alternative estimates over data-sparse areas globally like Africa and central Asia.

(39) Line 457: Another minor typo "is still challenged".

Response: Corrected as suggested.

(40) Line 463: It would be very beneficial to the community to share the data used in the manuscript figures, as well as the DATASET and bias-corrected CMIP6 members. This will have a positive impact on the study itself, as it will improve its reproducibility.

Response: We have made the data used in all the figures of the manuscript as well as the processed datasets and bias-corrected CMIP6 members publicly available via the Zenodo platform, which we will provide towards the later stages of the review.