The Annual Cycle of Terrestrial Water Storage Anomalies in CMIP6 Models Evaluated against GRACE Data

REN-JIE WU,^a MIN-HUI LO,^a AND BRIDGET R. SCANLON^b

^a *Department of Atmospheric Sciences, National Taiwan University, Taipei City, Taiwan*

^b *Bureau of Economic Geology, Jackson School of Geosciences, The University of Texas at Austin, Austin, Texas*

(Manuscript received 16 January 2021, in final form 5 July 2021)

ABSTRACT: The terrestrial water storage anomaly (TWSA) is a critical component of the global water cycle where improved spatiotemporal dynamics would enhance exploration of weather- and climate-linked processes. Thus, correctly simulating TWSA is essential not only for water-resource management but also for assessing feedbacks to climate through land–atmosphere interactions. Here we evaluate simulated TWSA from 25 climate models (from phase 6 of the Climate Model Intercomparison Project) through comparison with TWSA from GRACE satellite data (2003–14) in 14 river basins globally and assess causes of discrepancies by examining precipitation (P), evapotranspiration (ET), and runoff (R_{off}) fluxes during recharge (increasing TWS) and discharge (decreasing TWS) cycles. Most models show consistent biases in seasonal amplitudes of TWS anomalies relative to GRACE output: higher modeled amplitudes in river basins in high northern latitudes and the Parana and Congo basins, and lower amplitudes in most midlatitude basins and other tropical basins. This TWSA systematic bias also exists in the previous CMIP5 simulations. Models overestimate P compared to observed P datasets in 7 out of 14 basins, which increases (decreases) seasonal storage amplitude relative to GRACE in the recharge (discharge) cycle. Overestimation (underestimation) of runoff is another common contributing factor in the discharge phase that increases (decreases) TWSA amplitudes relative to GRACE in five river basins. The results provide a comprehensive assessment of the reliability of the simulated annual range in TWSA through comparison with GRACE data that can be used to guide future model development.

KEYWORDS: Hydrologic cycle; Precipitation; Runoff; Water masses/storage; Evapotranspiration; Satellite observations

1. Introduction

Land or terrestrial water storage (TWS) is a fundamental component of the hydrological cycle, and its variations can play an essential role in the climate system. Changes in TWS directly influence food and water security and can also lead to increased risk of floods and droughts (Reager et al. 2014; Thomas et al. 2014). Increases in interannual variability in soil moisture storage in the future were found over most of the high latitudes based on analysis of the representative concentration pathway (RCP) 8.5 projections performed as part of phase 5 of the Coupled Model Intercomparison Project (CMIP5) (Dirmeyer et al. 2013). Wu et al. (2015) also found increases in the annual range [defined as the difference between the maximum and minimum TWS anomaly (TWSA) in the mean seasonal cycle] of TWS in climate projections, mostly due to changes in soil moisture storage. Changes in subsurface water storage can affect the amount of stream baseflow from groundwater discharge (e.g., Lo et al. 2008). Furthermore, Milly and Dunne (2016) indicated that freshwater availability further decreases in currently water-stressed regions as climate changes.

Besides, changing TWS can also affect global atmospheric circulation (Milly and Dunne 1994). Changes in TWS contain information on previous climate forcings, and feedback from

these anomalies can affect subsequent seasons, as shown in the studies by Lo and Famiglietti (2011) and Wang et al. (2018), who found changes in precipitation P globally and regionally when considering groundwater information in climate models. The mid- to high-latitude river basins have a higher forecast skill in TWS decadal prediction, which is critical for freshwater management (Zhu et al. 2019). Moreover, realistically simulating TWS variability can improve climate predictions at seasonal to interannual time scales through land–atmosphere interactions.

TWSAs from the Gravity Recovery and Climate Experiment (GRACE) satellite mission since the launch of these satellites in 2002 (Tapley et al. 2004; Wahr et al. 2004) provide an excellent opportunity to evaluate and constrain model simulated TWSAs based on Earth's monthly gravity field. TWSAs include changes in the sum of snow, ice, surface water, soil moisture, and groundwater storage (Rodell and Famiglietti 1999). Monthly gravity variations obtained from GRACE provide information on changes in the spatiotemporal distribution of TWSA. Several studies have explored the accuracy and limitations of estimated GRACE basin-scale TWSA (e.g., Rodell and Famiglietti 1999; Swenson and Wahr 2002). In addition, GRACE data have been used to constrain model simulated TWSAs at monthly time scales (e.g., Lo et al. 2010; Werth et al. 2009; Zaitchik et al. 2008) to evaluate simulated TWSA seasonal amplitudes (Swenson and Milly 2006), to explore sea level changes at decadal time scales (Reager et al. 2016), and in many hydrological applications, such as interannual drought monitoring (Houborg et al. 2012; Thomas et al. 2014), flood prediction (Reager and Famiglietti 2009), and

Denotes content that is immediately available upon publication as open access.

Corresponding author: Min-Hui Lo, minhuilo@ntu.edu.tw

DOI: 10.1175/JCLI-D-21-0021.1

© 2021 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

agricultural irrigation water estimates (Anderson et al. 2012, 2015), as well as assessing long-term trends in global freshwater availability (Rodell et al. 2018). Scanlon et al. (2018) further indicate that opposite trends were found between modeled and GRACE derived trends in TWSA based on global averages: decreasing from models relative to increasing from GRACE between 2002 and 2014.

The uncertainties in monthly GRACE TWSAs are estimated to be ~ 1 cm equivalent water height at a horizontal resolution of $3^\circ \times 3^\circ$ (~ 300 km at the equator) (Wiese et al. 2016, 2018). GRACE satellites, therefore, offer unprecedented opportunities to explore global-scale TWSAs. Studies have shown that phase and amplitude biases exist in most model simulated TWSAs at seasonal, annual, and decadal time scales when compared to GRACE data (e.g., Döll et al. 2014; Freedman et al. 2014; Swenson and Milly 2006; Zeng et al. 2007; Zhang et al. 2017; Scanlon et al. 2019). These biases may be caused by P distribution variability, river contributions, and land process parameterizations. However, when analyzing the seasonal cycle in TWSA, it is useful to separately examine recharge and discharge periods, rather than annual means (Wu et al. 2015). The “recharge period” refers to the period when TWSA increases during a year whereas the “discharge period” refers to when TWSA decreases. For example, higher simulated versus observed P in the recharge period would lead to increased annual range in TWSA. Alternatively, during the discharge period, higher simulated versus observed P would maintain and replenish water storage, reducing the annual range in TWSA relative to GRACE.

Many studies have evaluated the hydrological fluxes in climate models for global river basins. For example, simulated precipitation was found to underestimate observed precipitation during the wet season over the Amazon (Yin et al. 2013). Significant differences in simulated versus observed P were found over the Congo Basin (Creese and Washington 2016). Almost all CMIP5 models under- and overestimated precipitation over the west and east of northern Eurasia, respectively (Hirota et al. 2016). Xue et al. (2017) showed that the CMIP5 ensemble model slightly overestimated P for each month relative to observed P while the seasonal variation in P is well captured in the Yangtze River basin. Also, several studies focused on evaluations of TWSA land water storage simulations. For example, Swenson and Milly (2006) used GRACE data to reveal climate model biases in seasonality of continental water storage from five CMIP3 models, especially underestimation of the seasonal range of TWSAs in low latitudes attributed to lack of a surface water component in the model and due to too early generation of annual maximum flow in the models. GRACE data were used to evaluate hydrological outputs for the Mississippi River basin from nine CMIP3 and nine CMIP5 climate models, in which increased agreement between the simulated annual cycle of TWSA relative to GRACE was shown for most CMIP5 models compared to earlier CMIP3 models (Freedman et al. 2014).

The objective of this study is to evaluate simulated TWSAs in CMIP6 models through comparison with GRACE TWSAs, focusing on the amplitude of the mean seasonal cycle (referred to as the annual range) using 14 major river basins distributed

globally across the five continents. Modeled underestimation of the annual range relative to GRACE may result from underestimation of inputs (i.e., P) or overestimation of outputs [i.e., evapotranspiration (ET) or runoff (R_{off})], and vice versa as indicated previously (Freedman et al. 2014; Swenson and Milly 2006). Therefore, we explore model simulated TWSAs during both recharge and discharge periods and quantify the monthly hydrological fluxes of the river basins (P , ET, R_{off}).

The main focus of this study is to understand the differences between the present hydrologic state in both observational and model data. We target the long-term mean seasonal cycle simulation in the models, instead of month-to-month comparisons. Another difference from previous studies by focusing on the water budgets during recharge and discharge cycles, rather than the annual mean. Furthermore, compared to the study of Swenson and Milly (2006), much longer GRACE records (2003–14) and more climate model simulations (25 in this study) are used to provide a more robust analysis of modeled versus GRACE global TWSAs. It is also essential to evaluate the land water storage in the climate model evolution and we will present the TWSA simulation efficiency from CMIP5 to CMIP6 in the discussion section.

2. Data and methods

a. Data

For observational data, TWSA is from GRACE satellite data (JPL GRACE MASCON RL06M; Wiese et al. 2018). Precipitation climatology data are from U.S. NOAA Global Precipitation Climatology Centre (GPCC; Schneider et al. 2011). Runoff climatology data are from observation-based global gridded runoff (GRUN; Ghiggi et al. 2019). ET is the satellite-based land ET [Global Land Evaporation Amsterdam Model (GLEAM); Martens et al. 2017; Miralles et al. 2011]. The time period for all of these observations is the same: 2003–14. To analyze the climatology of the land water storage cycle rather than annual variability, the seasonal cycle was calculated from all of the observational monthly data. This data processing provides consensus in observational data and represents current climate states at a regional scale.

For model simulations, we use outputs from a total of 25 CMIP6 historical models from 2003 to 2014 (Eyring et al. 2016; Taylor et al. 2017) and select the TWS variables [sum of snow water equivalent and soil moisture in all layers, and subtract the long-term (2003–14) means] as well as flux terms in the terrestrial water budget: P , ET, and R_{off} (Table 1). Both observational and model data are regridded to a 1° resolution.

b. Study basins

We select the nine largest river basins globally by drainage area for analysis plus five large river basins from the major continents, Africa, Australia, Asia, Europe, and North America (Fig. 1). These 14 basins (located from 40°S to 60°N) cover a wide range of climates. We focus on the annual range of TWSA and related hydrological fluxes over these river basins. The 14 basins are separated into three categories: tropical basins, including the Amazon, Congo, Ganges, Niger, and Nile basins;

TABLE 1. CMIP6 historical coupled models list. Expansions for most acronyms are available online at <http://www.ametsoc.org/PubsAcronymList>.

	Model ID	Institution
1	ACCESS-CM2	CSIRO-ARCCSS
2	ACCESS-EM1-5	CSIRO
3	AWI-ESM-1-1-LR	Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research
4	BCC-CSM2-MR	Beijing Climate Center
5	BCC-ESM1	Beijing Climate Center
6	CanESM5	Canadian Centre for Climate Modeling and Analysis
7	CESM2	National Center for Atmospheric Research
8	CESM2-WACCM	National Center for Atmospheric Research
9	CNRM-CM6-1	CNRM, CERFACS
10	CNRM-ESM2-1	CNRM, CERFACS
11	GFDL-CM4	National Oceanic and Atmospheric Administration, Geophysical Fluid Dynamics Laboratory
12	GISS-E2-1-G	Goddard Institute for Space Studies
13	GISS-E2-1-H	Goddard Institute for Space Studies
14	IPSL-CM6A-LR	Institut Pierre Simon Laplace
15	MIROC6	JAMSTEC, AORI, NIES, and R-CCS
16	MIROC-ES2L	JAMSTEC, AORI, NIES, and R-CCS
17	MPI-ESM-1-2-HAM	Max Planck Institute for Meteorology
18	MPI-ESM1-2-HR	Max Planck Institute for Meteorology
19	MPI-ESM1-2-LR	Max Planck Institute for Meteorology
20	MPI-ESM2-0	Max Planck Institute for Meteorology
21	NorESM2-LM	Norwegian Climate Centre
22	NorESM2-MM	Norwegian Climate Centre
23	SAM0-UNICON	Seoul National University
24	TaiESM1	Sinica, Taiwan
25	UKESM1-0-LL	Met Office Hadley Centre

midlatitude basins, including the Danube, Mississippi, Murray, Parana, and Yangtze basins; and northern high-latitude basins, including the Lena, Mackenzie, Ob, and Yenisei basins.

c. Methods

1) TAYLOR DIAGRAM

Taylor diagrams are typically used to compare simulations with observations based on a statistical summary (Taylor 2001). We use Taylor diagrams to determine how well model datasets match a reference dataset (i.e., GRACE TWSA annual cycle, including phase and amplitude). Several statistical measures of the differences between two datasets are shown in polar coordinates (r and θ). The similarity between two datasets is quantified in terms of their correlation r ($r = \cos\theta$), their normalized root-mean-square error [RMSE; represented by the distance to (1, 0) on the x axis in the Taylor diagram], and the amplitude of their variations, represented by their normalized standard deviation ratio ($\sigma_{\text{CMIP5}}/\sigma_{\text{obs}}$). Normalized standard deviations exceeding 1.0 indicate model variability larger than observations.

2) WATER BUDGET EQUATION (P , ET , R_{OFF})

The annual range is defined as the difference between the maximum and minimum TWSA in the mean seasonal cycle. The mean TWSA seasonal cycle can then be separated into two parts: the recharge period extending from the time of the TWSA minimum to the time of the TWS maximum (increasing

TWSA) and the opposite for the discharge period (TWSA maximum to minimum). To explore uncertainties in the annual range in TWSA, we used the following water budget equation:

$$dS_{\text{rech}} = \int_{\min_{\text{mo}}}^{\max_{\text{mo}}} (P - ET - R_{\text{off}}) dt;$$

$$dS_{\text{disch}} = \int_{\max_{\text{mo}}}^{\min_{\text{mo}}} (P - ET - R_{\text{off}}) dt, \quad (1)$$

where S is TWSA, dS is TWSA range, and rech (disch) denotes the recharge (discharge) period of the mean seasonal cycle. For example, in the Amazon, the discharge period includes the months from April through October, while the recharge period is from October through April (Fig. 2). During the discharge phase, water storage declines from April to October by ~ 400 mm. Because GRACE data are considered as a monthly average, we sum the precipitation (cyan bars in Fig. 2) from May to September (including 50% of the precipitation in April and October, a central differencing approach) as the discharge period accumulated P (ΣP). The same processing is performed on ET and R_{off} to obtain the discharge period estimate of accumulated ET (ΣET) and accumulated R_{off} (ΣR_{off}). During the recharge phase, we use annual total P minus the discharge period total P to obtain the recharge period total P . These terms in the water budget equation help us understand how biases in the annual range in TWSA are related to biases in particular hydrological fluxes in the recharge and discharge periods.

FIG. 1. Location of the 14 river basins and their TWSA seasonal cycles examined in this study. Basins colored blue (red) have modeled annual range in TWSA less (greater) than that from GRACE in the majority of CMIP6 (>16 out of 25) models. Black solid lines denote the annual cycles computed using GRACE data; the blue lines were derived from the 25 models. The values in the corner of the y axis in each plot denote the minimum and maximum range of the y axis. The numbers in parentheses (X, Y) next to the river IDs represent the number of models with smaller/larger TWSA ranges relative to GRACE, respectively.

According to the water budget equation: $dS = \int (P - ET - R_{\text{off}}) dt$, biases in the annual TWSA range may be partitioned into integrations of flux terms (P , ET , or R_{off}) during the recharge and discharge periods. For example, during the recharge period ($dS > 0$), lower P , higher ET , or higher R_{off} relative to observations can separately or in combination result in a lower TWSA peak, which can result in a lower annual range in simulated TWSA compared to that from GRACE. Alternatively, during the discharge period ($dS < 0$), a lower modeled annual range in TWSA relative to GRACE may result from modeled higher P , lower ET , or lower R_{off} than observational fluxes.

We calculate ΣP , ΣET , and ΣR_{off} during both the recharge and discharge periods for each model and compare the fluxes with the observed fluxes. If ΣP , ΣET , and ΣR_{off} can explain the biased TWSA annual range based on Eq. (1), we refer to the terms ΣP , ΣET , or ΣR_{off} as “contributing factors” in this study. Based on TWSA from each model during the discharge and recharge periods, we represent the interannual variability in both TWSA phases by one standard deviation (STD) of observational ΣP , ΣET , and ΣR_{off} value from 2003 to 2014. We highlight in the figure when the contributing factor flux is higher or lower than the mean ± 1 STD range of the corresponding observational value. Furthermore, not all CMIP6 simulations have a lower annual range than that in GRACE. We focus on models with the dominant bias in the annual range in TWSA in each basin (i.e., >16 out of 25 models). The next step is to find the common contributing factors in each river basin with the same threshold. If more than 16 models show the same contributing factors in a river basin, we consider these contributing factors as “dominant contributing factors.” The final step is to assess the significance of dominant contributing

factors, which is represented by the ratio of red dot models to total contributing factor models. If the ratio exceeds 50%, 66%, or 75%, we mark with 1, 2, or 3 asterisks, respectively, in the oblong in the dominant contributing factor map. For

FIG. 2. Water budget schematic for the Amazon River basin. The black solid line is the climatological monthly values of GRACE TWSA for the period 2003–14. The cyan bars are GPCC precipitation data. The yellow bars are ET from GLEAM. The green bars are runoff from GRUN. The blue arrow along the x axis denotes the discharge period (TWSA decreasing); the dark green arrow refers to the recharge period (TWSA increasing).

FIG. 3. Scatterplots of modeled (red dots) and observational (blue and green dots) water budget data for the 14 river basins. The x axis is the TWSA range (dS), and the y axis is the summation of $P-ET-R_{\text{off}}$ during TWSA recharge and discharge periods. The solid 1:1 line represents the water budget conservation line ($y = x$), meaning the summation of $P-ET-R_{\text{off}}$ is equal to the dS . Two gray lines are the 50% nonclosure boundary. The vertical dashed line is the TWSA discharge ($dS < 0$) and recharge ($dS > 0$) period boundary. Green and blue dots are the mean of observational data results for the TWSA recharge and discharge period, respectively.

example, Fig. 1 shows that 21 (4) models have lower (higher) TWSA annual ranges than that in GRACE in the Nile River basin. Thus, we consider the simulated TWSA annual range is underestimated in the Nile and only explore the water budget of those 21 models. Then, we assign the dominant contributing factors (>16 out of 25) of those 21 models with lower TWSA annual ranges as the dominant contributing factor. There are

17 models with ΣP higher than observations in those 21 models with 15 of the 17 models exceeding the 1 STD range. As a result, ΣP is the dominant contributing factor over the Nile River basin in TWSA discharge season with three asterisks (15/17, $>75\%$) in the dominant contributing map. Knowing these critical contributing factors from the water budget in each river basin during recharge and discharge periods can help us better

FIG. 4. TWSA seasonal cycle Taylor diagrams for the 14 river basins: (a) tropical basins, (b) midlatitude basins, and (c) northern high-latitude basins. The black straight dashed lines are the boundary correlations (0.6 and 0.9) to GRACE. The curve dashed line is the normalized standard deviation (the standard deviation of the seasonal cycle of the model divided by that of the GRACE observations). The GRACE data are shown as “REF” in the figure. Also shown are boxplots for (d) the centered pattern of root-mean-square error (RMSE) and (e) the ratio of the simulated TWSA annual range to that from GRACE. The boxplots show the interquartile range with a rectangular box; the red bar, which may be inside or outside of the interquartile range, represents the average for 25 models. The basin numbers are shown in Fig. 1.

understand the source of TWSA biases in the land hydrological cycle in current climate models.

3. Results

a. Water budget conservation

Figure 3 is a scatterplot of water budget closure and all of the data are analyzed in the 2003 to 2014 seasonal cycle. For most CMIP6 models, the water budget is near or on the closure line for both TWSA recharge and discharge seasons. Although there are nine points out of 50% of nonclosure boundary in the observations, the observational data are reasonably conserved in either TWSA recharge or discharge phase for all 14 river basins. The imbalance of observational data may be due to different measurement methods for P , ET , R_{off} , and TWSA data, including satellite, station location, or model reanalysis among those observational data.

b. Modeled total water storage anomalies

In most river basins, CMIP6 simulations have phases that are similar to those from GRACE with the exception of basins 11–14 in the northern high latitudes (Fig. 1). Correlation

coefficients between simulated and GRACE mean seasonal TWSAs mostly exceed 0.6 (Figs. 4a–c) and pass a 5% significance t test in most river basins, except for a few models in the Murray and Yangtze River basins. Tropical basins show high correlation coefficients most likely due to the larger annual range, while most midlatitude basins have variable correlation coefficients. Simulated annual TWSA ranges, on the other hand, are variable across different latitudes, usually higher in high northern latitude basins and lower in most midlatitude and tropical basins when compared to GRACE data, except for the Parana and Congo basins (Figs. 1 and 4e). However, modeled TWSA biases relative to GRACE in the Murray and Danube River basins are statistically insignificant, with no majority (>16) in models having the same TWSA bias.

Considering both the recharge and discharge phases as well as the annual range, the box plot of centered pattern RMSE (Fig. 4d) shows a lower centered pattern for RMSE values in the Danube and Mississippi River basins. Red bars in Figs. 4d and 4e are the mean states of all 25 models in each river basin. Red bars above the interquartile box mean that the data are negatively skewed. The mean RMSE trends toward positive extreme RMSE in three high-latitude river basins, and this overestimation of the TWSA range in higher latitudes

FIG. 5. Map showing dominant contributing factors to TWSA annual ranges. Basins colored red have modeled annual ranges that exceed observed ranges whereas basins colored in blue show modeled TWSA ranges less than observed. There are six types of dominant contributing factors, which are marked with shading oblong(s) next to their river basin mask: ΣP in both TWSA recharge and discharge is in shallow and deep blue color bar, respectively; ΣET is in the yellow series color bar; ΣR_{off} is in the green series color bar. The number of asterisks (1, 2, or 3) in the oblong(s) indicates that there are over 50%, 66%, or 75%, respectively, of CMIP6 models with more or less than the mean ± 1 STD range of observational accumulated flux among the models having contributing factors.

generally results in larger RMSEs than in other basins. Similar biases were found by Swenson and Milly (2006) for some of the CMIP3 models.

Not all of the models show the same bias tendency in the TWSA range in each river basin (Fig. 4c); thus, we target the models with the same majority TWSA range bias (from the water budget perspective as shown in the method section) over 14 river basins for those models with the same majority TWSA range bias (see Fig. 1 values in parentheses). Moreover, six blue (six red) shaded river basin masks have a smaller (larger) TWSA range than GRACE data in at least 17 models.

1) UNDERESTIMATED ANNUAL TWSA RANGE: SIX RIVER BASINS

A total of 6 out of 14 basins have lower simulated TWSA annual ranges than GRACE, mostly in tropical and midlatitude regions (basins 1, 3, 4, 5, 7, and 8; Fig. 5). For simulations with lower TWSA ranges relative to GRACE, ΣP should be lower (higher) than observational ΣP in the TWSA recharge (discharge) period. Alternatively, ΣET and ΣR_{off} should be higher (lower) than observational data during the recharge (discharge) period. The dominant contributing factors are the higher R_{off} during the recharge period, found in two out of the six basins (Table 2), followed by higher P during the discharge period (again two out of six basins).

2) OVERESTIMATED ANNUAL TWSA RANGE: SIX RIVER BASINS

A total of 6 out of 14 river basins have simulated TWSA annual ranges exceeding those from GRACE (Fig. 1). These basins are found in the four northern high latitudes (above 60°N), and the Parana and the Congo. Most basins (five out of

six, all except the Parana) have higher ΣP than observations during the TWSA recharge period, which is the main contributing factor to the TWSA bias. The ΣR_{off} is another dominant contributing factor in a larger TWSA annual range (Fig. 5) that occurs in all high-latitude river basins in the Northern Hemisphere.

c. Dominant contributing factor

According to the map of the dominant contributing factors (Fig. 5), accumulated precipitation bias plays an essential role in the TWSA difference between models and GRACE in many river basins. Models show higher annual P than observational data in seven river basins (7/14), including river basins with five larger and two smaller simulated TWSA ranges relative to GRACE. Those five larger TWSA basins with more P than observation are in the recharge period of TWSA. On the other hand, the two smaller TWSA river basins have more P in the discharge period. The Ganges is the only basin with less P as the dominant contributing factor in the TWSA recharge season, which would result in TWSA less than GRACE. Another critical common dominant contributing factor is more accumulated R_{off} in six river basins (6/14); four of those six river basins overestimate R_{off} during the discharge period, which would result in a higher TWSA range relative to GRACE. The ET component was not found to be a dominant contributing factor in most river basins (only 2/14). These dominant contributing factors are not only qualitative comparing to the target; most of them (except for the Ob and Ganges) deviate from 1 STD of the observed interannual variations.

The ET contributing factor appears in several basins with significant differences from observations, such as the Ganges, Nile, and Yangtze. More than half of the models have recharge

TABLE 2. Dominant contributing factors (i.e., show consistency in contributing factors over 2/3 of the 25 climate models) in the 14 river basins resulting in lower and higher TWSA annual ranges relative to those from GRACE. Numbers refer to river IDs in Fig. 1.

Dominant contributing factors	Lower TWSA	Higher TWSA
Discharge— ΣP	5, 8	—
Discharge— ΣET	1	—
Discharge— ΣR_{off}	1	11, 12, 13, 14
Recharge— ΣP	3	2, 11, 12, 13, 14
Recharge— ΣET	7	—
Recharge— ΣR_{off}	3, 8	—

ET contributing factors, but fewer than two-thirds of the 25 CMIP6 models have the same contributing factor. The contributing factors for the individual models of all river basins are shown in Figs. 6 and 7. The ratio of the models with significant changes in contributing factors to the total number of models with contributing factors exceeds 75% in 12 of the 14 river basins, except for the Ob (66%) and the Ganges (62%). Note that the Danube and Murray are the two river basins with unclear TWSA range biases, and the Parana and Niger do not

have dominant contributing factors. This is most likely related to highly variable simulated P , ET , and R_{off} over these basins that do not result in systematic differences in modeled versus GRACE TWSA during both recharge and discharge periods; therefore, it is challenging to determine dominant contributing factors on those four basins.

4. Discussion

This study calculates seasonal cycles in TWSA from 25 CMIP6 models, comparing these models with GRACE MASCON data from 2003 to 2014 in 14 river basins globally. The MASCON seasonal cycle is highly correlated to the other GRACE ensemble datasets with relatively small root-mean-square differences compared to the mean state. Because the MASCON product is the latest version with the longest record relative to the other three GRACE products, we use GRACE MASCON data as the observational TWSA. The TWSA seasonal cycle in the climate models contains systematic biases among these river basins, which is highly related to latitude. While this study targets the mean seasonal cycle evaluations, we also use CMIP5 and CMIP6 historical runs compared to GRACE with overlap period of 2003–05 (Fig. 8). The biases still exist in the

FIG. 6. The bar chart of 25 models' contributing factors for the tropical river basins with six river basins having simulated TWSA range less than GRACE: (a) Amazon, (b) Ganges, (c) Mississippi, (d) Niger, (e) Nile, and (f) Yangtze. The model names and the corresponding institutions are shown in Table 1. The contributing factor of precipitation in both TWSA discharge and recharge is in light and dark blue color bars, respectively; the contributing factor of evapotranspiration is in the yellow series color bar; the contributing factor of runoff is in the green series color bar. The red dots in the contributing factors mean the simulated accumulated flux is more or less than the mean ± 1 STD of annual accumulated observational data from 2003 to 2014 based on each model's TWSA period.

FIG. 7. As in Fig. 6, but with basins having larger TWSA range biases: (a) Congo, (b) Lena, (c) Mackenzie, (d) Ob, (e) Parana, and (f) Yenisei river basins.

CMIP6 models, which are underestimated in mid- to low latitudes and overestimated in higher latitudes in most river basins. For example, in the Amazon and Ganges basins, CMIP6 models have a larger TWSA range than CMIP5 models but still underestimate the TWSA range compared to GRACE. By contrast, the seasonal cycle phase in CMIP6 models is improved in many river basins with a correlation coefficient of around 0.9 relative to GRACE data. This result provides information on the limitation of the TWSA improvement in the updated CMIP6 climate models. This latitudinal dependence of TWSA bias is also found by using different periods for historical climate models, suggesting that the TWSA seasonal cycle biases are essentially stationary.

Note that models 1 and 2 are from the same modeling group (Australia CSIRO-ARCCSS Institute): one is the Earth system model (ACCESS-ESM1.5) and the other is the coupled model (ACCESS-CM2). The major difference between the two models is that the carbon cycle mechanism is interactive in ACCESS-ESM1.5 but not in ACCESS-CM2 (Ziehn et al. 2020; Bi et al. 2020). According to Bi et al. (2020), the leaf area index (LAI) seasonal cycle is overestimated in ACCESS-ESM1.5, corresponding to higher simulated ET. This feature can be seen in the results among the Mackenzie, Lena, and Yenisei basins in this study (model 2; dark yellow bar in Fig. 7). On the other hand, ACCESS-CM2 (without an interactive carbon cycle) does not have the same ET biases as ACCESS-ESM1.5 in many river basins, indicating that the uncertainty in the ET simulation in the model may result from different geochemistry schemes.

This main result from this study helps us understand which flux terms (sources or sinks) would lead to the TWSA magnitude bias. Crowley et al. (2006) indicated that P roughly contributes 3 times as much to the TWSA peak in the Congo rainy season. The model simulated P is also slightly larger than the observation at the interannual scale. We use longer and more comprehensive data to find the contributing factors in the recharge and discharge periods, which is focused on seasonality instead of annual flux terms. Furthermore, from the water budget closure scatterplot (Fig. 3), the well-conserved situation is mostly shown in CMIP6 and observational data in both TWSA recharge and discharge periods. There are nine points out of 50% conserved line for the observational data, which are over the Amazon, Congo, Yenisei, and Lena basins in the TWSA recharge period and over the Ganges, Mackenzie, Mississippi, Parana, and Yangtze in the TWSA discharge period. We have done a sensitivity test of the biased observational fluxes (P , ET , or R_{off}) on the results of the contributing factors for all of the basins. To explore the maximum possible influence from the nonclosure observational water budget on the selected contributing factor, we assume the bias [difference between TWSA range and $\Sigma(P-ET-R_{\text{off}})$ in both phases] is derived from one of the fluxes (P , ET , or R_{off}). This nonclosure bias [$\text{TWSA} - \Sigma(P-ET-R_{\text{off}})$ in both phases] is then added to each P , ET , or R_{off} , and weighted by their seasonal cycle. Then, we reselect the dominant contributing factors under those three extreme situations. The result shows that the P and R_{off} dominant contributing factor is robust in the climate models even when we add all the nonclosure bias to one of the

FIG. 8. Taylor diagram of modeled TWSA from 19 CMIP5 models (red dot group) and 25 CMIP6 models (blue cross group) in 14 river basins compared to GRACE data. All the data are based on the averaged seasonal cycle from 2003 to 2005.

components. However, under the extreme situation, the uncertainty in ET observational data would make the ET contributing factor changes significant in some basins, such as the Amazon, Ganges, Niger, Mackenzie, Yangtze, and Yenisei basins.

The groundwater component is also an important variable that could affect not only the amount of water storage but also the land hydrological processes. Rahman et al. (2019) indicated that adding the groundwater flow boundary in the model improves the soil moisture profile in different soil types, and shows a close accumulated discharge rate to the fully integrated three-dimensional hydrological model. Most CMIP6

models do not consider the groundwater component in the corresponding land surface models, and all of the CMIP members only offer soil moisture and snow water equivalent, which is regarded as TWSA in this study. Thus, we further conduct an offline CESM-CLM and a coupled TaiESM-CLM simulation, in which a simple lumped groundwater component was included. We then compare the simulated TWSA seasonal cycle with and without groundwater for 14 river basins from 2003 to 2005. The results show that the groundwater component contributes small variations to the TWSA annual range and only changes the TWSA phase

FIG. 9. Taylor diagrams of modeled TWSA compared to GRACE data. Numbers 1 and 3 indicate the CLM4 TWSA offline runs with and without groundwater component, respectively. Number 2 indicates the CLM4 TWSA offline simulation without the river contribution component. Numbers 4 and 5 are the TaiESM TWSA simulations with and without groundwater component, respectively. All the data are based on the averaged seasonal cycle from 2003 to 2005.

slightly in most river basins. The systematic TWSA range bias for each river basin still remains in the model after considering the groundwater component (Fig. 9). Yuan and Zhu (2018) highlighted that groundwater would retain multiyear memory in TWS predictability in most land areas globally, which may affect the TWSA seasonal cycle in the land model. Besides the groundwater component, river water contribution is another source of TWSA (Kim et al. 2009). For example, in the offline CLM simulations, removing the river water component lowers the TWSA ranges in the Amazon and Congo basins.

5. Conclusions

Comparison of simulated versus GRACE TWSAs in 14 river basins globally shows systematic biases in TWSA annual ranges for most of the 25 models relative to GRACE. These biases are generally latitudinally dependent and have appeared since CMIP5 models. Six of the basins have modeled TWSA annual ranges that are lower than those from GRACE, mainly located in the tropics and midlatitudes. The main contributing factors to the lower modeled TWSA are higher $\sum R_{\text{off}}$ relative to observations during the recharge period (two out of these six basins) and higher $\sum P$ during the discharge period (two out of these six basins). The other six basins have simulated larger TWSA than GRACE, and are in the Northern Hemisphere high-latitude basins and the Congo. The dominant contributing factor to higher modeled TWSA relative to GRACE is the higher modeled P than the observations during the recharge period in northern high latitudes. Furthermore, in the discharge season from spring to summer overestimated R_{off} could be detected in most of the CMIP6 models. The contributing factors analysis shows that there is some consistency in land process bias in the current climate models. In this study, the water budget analysis is conducted and the key contributing factors that could explain the TWSA bias in these models are determined. Furthermore, the dominant contributing factors during both recharge and discharge periods highlight areas for

potential improvements in modeled TWSA. This approach can also be applied to different future scenarios or spatial resolutions in related research.

Acknowledgments. This study was supported by the grants of MOST 106-2111-M-001-005 and MOST 110-2628-M-002-004-MY4 to National Taiwan University. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. For CMIP, the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. GPCC precipitation data are provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA. GRACE land data are available and supported by the NASA MEaSUREs Program. We acknowledge Global Land Evaporation Amsterdam Model (GLEAM) data and the Global Runoff (GRUN) data provided from the website.

Data availability statement. The data used in this study can be obtained as follows: 1) CMIP6: <https://esgf-node.llnl.gov/projects/cmip6/>; 2) GRACE: <http://grace.jpl.nasa.gov/>; 3) GPCC: ftp://ftp.cgd.noaa.gov/Datasets/gpcc/full_v2018/precip.mon.total.v2018.nc; 4) GLEAM: <https://www.gleam.eu/>; 5) GRUN: <https://doi.org/10.6084/m9.figshare.9228176>. All the data processing and figures code can be downloaded in Github (https://github.com/wurjicnip/CMIP6_Figures.git).

REFERENCES

- Anderson, R. G., M.-H. Lo, and J. S. Famiglietti, 2012: Assessing surface water consumption using remotely-sensed groundwater, evapotranspiration, and precipitation. *Geophys. Res. Lett.*, **39**, L16401, <https://doi.org/10.1029/2012GL052400>.
- , —, S. Swenson, J. S. Famiglietti, Q. Tang, T. H. Skaggs, Y.-H. Lin, and R.-J. Wu, 2015: Using satellite-based estimates of evapotranspiration and groundwater changes to determine

- anthropogenic water fluxes in land surface models. *Geosci. Model Dev.*, **8**, 3021–3031, <https://doi.org/10.5194/gmd-8-3021-2015>.
- Bi, D., and Coauthors, 2020: Configuration and spin-up of ACCESS-CM2, the new generation Australian Community Climate and Earth System Simulator Coupled Model. *J. South. Hemisphere Earth Syst. Sci.*, **70**, 225, <https://doi.org/10.1071/ES19040>.
- Creese, A., and R. Washington, 2016: Using qflux to constrain modeled Congo Basin rainfall in the CMIP5 ensemble. *J. Geophys. Res.*, **121**, 13 415–13 442, <https://doi.org/10.1002/2016JD025596>.
- Crowley, J. W., J. X. Mitrovica, R. C. Bailey, M. E. Tamisiea, and J. L. Davis, 2006: Land water storage within the Congo Basin inferred from GRACE satellite gravity data. *Geophys. Res. Lett.*, **33**, L19402, <https://doi.org/10.1029/2006GL027070>.
- Dirmeyer, P. A., Y. Jin, B. Singh, and X. Yan, 2013: Trends in land–atmosphere interactions from CMIP5 simulations. *J. Hydrometeor.*, **14**, 829–849, <https://doi.org/10.1175/JHM-D-12-0107.1>.
- Döll, P., M. Fritsche, A. Eicker, and H. Müller Schmied, 2014: Seasonal water storage variations as impacted by water abstractions: Comparing the output of a global hydrological model with GRACE and GPS observations. *Surv. Geophys.*, **35**, 1311–1331, <https://doi.org/10.1007/s10712-014-9282-2>.
- Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, 2016: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958, <https://doi.org/10.5194/gmd-9-1937-2016>.
- Freedman, F. R., K. L. Pitts, and A. F. C. Bridger, 2014: Evaluation of CMIP climate model hydrological output for the Mississippi River Basin using GRACE satellite observations. *J. Hydrol.*, **519**, 3566–3577, <https://doi.org/10.1016/j.jhydrol.2014.10.036>.
- Ghiggi, G., V. Humphrey, S. I. Seneviratne, and L. Gudmundsson, 2019: GRUN: An observation-based global gridded runoff dataset from 1902 to 2014. *Earth Syst. Sci. Data*, **11**, 1655–1674, <https://doi.org/10.5194/essd-11-1655-2019>.
- Hirota, N., Y. N. Takayabu, and A. Hamada, 2016: Reproducibility of summer precipitation over northern Eurasia in CMIP5 multiclimate models. *J. Climate*, **29**, 3317–3337, <https://doi.org/10.1175/JCLI-D-15-0480.1>.
- Houborg, R., M. Rodell, B. Li, R. Reichle, and B. F. Zaitchik, 2012: Drought indicators based on model-assimilated Gravity Recovery and Climate Experiment (GRACE) terrestrial water storage observations. *Water Resour. Res.*, **48**, W07525, <https://doi.org/10.1029/2011WR011291>.
- Kim, H., P. J.-F. Yeh, T. Oki, and S. Kanae, 2009: Role of rivers in the seasonal variations of terrestrial water storage over global basins. *Geophys. Res. Lett.*, **36**, L17402, <https://doi.org/10.1029/2009GL039006>.
- Lo, M.-H., and J. S. Famiglietti, 2011: Precipitation response to land subsurface hydrologic processes in atmospheric general circulation model simulations. *J. Geophys. Res.*, **116**, D05107, <https://doi.org/10.1029/2010JD015134>.
- , P. J. F. Yeh, and J. S. Famiglietti, 2008: Constraining water table depth simulations in a land surface model using estimated baseflow. *Adv. Water Resour.*, **31**, 1552–1564, <https://doi.org/10.1016/j.advwatres.2008.06.007>.
- , J. S. Famiglietti, P. J. F. Yeh, and T. H. Syed, 2010: Improving parameter estimation and water table depth simulation in a land surface model using GRACE water storage and estimated base flow data. *Water Resour. Res.*, **46**, W05517, <https://doi.org/10.1029/2009WR007855>.
- Martens, B., and Coauthors, 2017: GLEAM v3: Satellite-based land evaporation and root-zone soil moisture. *Geosci. Model Dev.*, **10**, 1903–1925, <https://doi.org/10.5194/gmd-10-1903-2017>.
- Milly, P. C. D., and K. A. Dunne, 1994: Sensitivity of the global water cycle to the water-holding capacity of land. *J. Climate*, **7**, 506–526, [https://doi.org/10.1175/1520-0442\(1994\)007<0506:SOTGWC>2.0.CO;2](https://doi.org/10.1175/1520-0442(1994)007<0506:SOTGWC>2.0.CO;2).
- , and —, 2016: Potential evapotranspiration and continental drying. *Nat. Climate Change*, **6**, 946–949, <https://doi.org/10.1038/nclimate3046>.
- Miralles, D. G., T. R. H. Holmes, R. A. M. de Jeu, J. H. Gash, A. G. C. A. Meesters, and A. J. Dolman, 2011: Global land-surface evaporation estimated from satellite-based observations. *Hydrol. Earth Syst. Sci.*, **15**, 453–469, <https://doi.org/10.5194/hess-15-453-2011>.
- Rahman, M., R. Rosolem, S. J. Kollet, and T. Wagener, 2019: Towards a computationally efficient free-surface groundwater flow boundary condition for large-scale hydrological modeling. *Adv. Water Resour.*, **123**, 225–233, <https://doi.org/10.1016/j.advwatres.2018.11.015>.
- Reager, J. T., and J. S. Famiglietti, 2009: Global terrestrial water storage capacity and flood potential using GRACE. *Geophys. Res. Lett.*, **36**, L23402, <https://doi.org/10.1029/2009GL040826>.
- , B. F. Thomas, and J. S. Famiglietti, 2014: River basin flood potential inferred using GRACE gravity observations at several months lead time. *Nat. Geosci.*, **7**, 588–592, <https://doi.org/10.1038/ngeo2203>.
- , A. S. Gardner, J. S. Famiglietti, D. N. Wiese, A. Eicker, and M.-H. Lo, 2016: A decade of sea level rise slowed by climate-driven hydrology. *Science*, **351**, 699–703, <https://doi.org/10.1126/science.aad8386>.
- Rodell, M., and J. S. Famiglietti, 1999: Detectability of variations in continental water storage from satellite observations of the time-dependent gravity field. *Water Resour. Res.*, **35**, 2705–2723, <https://doi.org/10.1029/1999WR900141>.
- , —, D. N. Wiese, J. T. Reager, H. K. Beaudoin, F. W. Landerer, and M. H. Lo, 2018: Emerging trends in global freshwater availability. *Nature*, **557**, 651–659, <https://doi.org/10.1038/s41586-018-0123-1>.
- Scanlon, B. R., and Coauthors, 2018: Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. *Proc. Natl. Acad. Sci. USA*, **115**, E1080–E1089, <https://doi.org/10.1073/pnas.1704665115>.
- , and Coauthors, 2019: Tracking seasonal fluctuations in land water storage using global models and GRACE satellites. *Geophys. Res. Lett.*, **46**, 5254–5264, <https://doi.org/10.1029/2018GL081836>.
- Schneider, U., A. Becker, P. Finger, A. Meyer-Christoffer, B. Rudolf, and M. Ziese, 2011: GPCC full data reanalysis version 6.0 at 1.0°: Monthly land-surface precipitation from rain-gauges built on GTS-based and historic data. Accessed 26 July 2018, https://doi.org/10.5676/DWD_GPCC/FD_M_V6_050.
- Swenson, S., and J. Wahr, 2002: Methods for inferring regional surface-mass anomalies from Gravity Recovery and Climate Experiment (GRACE) measurements of time-variable gravity. *J. Geophys. Res.*, **107**, 2193, <https://doi.org/10.1029/2001JB000576>.
- , and P. C. D. Milly, 2006: Climate model biases in seasonality of continental water storage revealed by satellite gravimetry. *Water Resour. Res.*, **42**, W03201, <https://doi.org/10.1029/2005WR004628>.

- Tapley, B., S. Bettapur, M. Watkins, and C. Reigber, 2004: The Gravity Recovery and Climate Experiment: Mission overview and early results. *Geophys. Res. Lett.*, **31**, L09607, <https://doi.org/10.1029/2004GL019920>.
- Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106**, 7183–7192, <https://doi.org/10.1029/2000JD900719>.
- , and Coauthors, 2017: CMIP6 global attributes, DRS, filenames, directory structure, and CV's. version v6.2.3 (4 April 2017), https://docs.google.com/document/d/1h0r8RZr_f3-8egBMMh7aqLwy3snpD6_MrDz1q8n5XUk.
- Thomas, A. C., J. T. Reager, J. S. Famiglietti, and M. Rodell, 2014: A GRACE-based water storage deficit approach for hydrological drought characterization. *Geophys. Res. Lett.*, **41**, 1537–1545, <https://doi.org/10.1002/2014GL059323>.
- Wahr, J., S. Swenson, V. Zlotnicki, and I. Velicogna, 2004: Time-variable gravity from GRACE: First results. *Geophys. Res. Lett.*, **31**, L11501, <https://doi.org/10.1029/2004GL019779>.
- Wang, F., A. Ducharme, F. Cheruy, M. H. Lo, and J. L. Grandpeix, 2018: Impact of a shallow groundwater table on the global water cycle in the IPSL land-atmosphere coupled model. *Climate Dyn.*, **50**, 3505–3522, <https://doi.org/10.1007/s00382-017-3820-9>.
- Werth, S., A. Güntner, S. Petrovic, and R. Schmidt, 2009: Integration of GRACE mass variations into a global hydrological model. *Earth Planet. Sci. Lett.*, **277**, 166–173, <https://doi.org/10.1016/j.epsl.2008.10.021>.
- Wiese, D. N., F. W. Landerer, and M. M. Watkins, 2016: Quantifying and reducing leakage errors in the JPL RL05M GRACE mascon solution. *Water Resour. Res.*, **52**, 7490–7502, <https://doi.org/10.1002/2016WR019344>.
- , D.-N. Yuan, C. Boening, F. W. Landerer, and M. M. Watkins, 2018: JPL GRACE mascon ocean, ice, and hydrology equivalent water height release 06 coastal resolution improvement (CRI) filtered version 1.0. PO.DAAC, accessed 1 October 2018, <https://doi.org/10.5067/TEMSC-3MJC6>.
- Wu, W.-Y., C.-W. Lan, M.-H. Lo, J. T. Reager, and J. S. Famiglietti, 2015: Increases in the annual range of soil water storage at northern middle and high latitudes under global warming. *Geophys. Res. Lett.*, **42**, 3903–3910, <https://doi.org/10.1002/2015GL064110>.
- Xue, L., B. Zhu, C. Yang, G. Wei, X. Meng, A. Long, and G. Yang, 2017: Study on the characteristics of future precipitation in response to external changes over arid and humid basins. *Sci. Rep.*, **7**, 15148, <https://doi.org/10.1038/s41598-017-15511-5>.
- Yin, L., R. Fu, E. Shevliakova, and R. E. Dickinson, 2013: How well can CMIP5 simulate precipitation and its controlling processes over tropical South America? *Climate Dyn.*, **41**, 3127–3143, <https://doi.org/10.1007/s00382-012-1582-y>.
- Yuan, X., and E. Zhu, 2018: A first look at decadal hydrological predictability by land surface ensemble simulations. *Geophys. Res. Lett.*, **45**, 2362–2369, <https://doi.org/10.1002/2018GL077211>.
- Zaitchik, B. F., M. Rodell, and R. Reichle, 2008: Assimilation of GRACE terrestrial water storage data into a land surface model: Results for the Mississippi River basin. *J. Hydrometeorol.*, **9**, 535–548, <https://doi.org/10.1175/2007JHM951.1>.
- Zeng, N., J.-H. Yoon, A. Mariotti, and S. Swenson, 2007: Variability of basin-scale terrestrial water storage from a PER water budget method: The Amazon and the Mississippi. *J. Climate*, **21**, 248–265, <https://doi.org/10.1175/2007JCLI1639.1>.
- Zhang, L., H. Dobslaw, T. Stacke, A. Güntner, R. Dill, and M. Thomas, 2017: Validation of terrestrial water storage variations as simulated by different global numerical models with GRACE satellite observations. *Hydrol. Earth Syst. Sci.*, **21**, 821–837, <https://doi.org/10.5194/hess-21-821-2017>.
- Zhu, E., X. Yuan, and A. Wood, 2019: Benchmark decadal forecast skill for terrestrial water storage estimated by an elasticity framework. *Nat. Commun.*, **10**, 1237, <https://doi.org/10.1038/s41467-019-09245-3>.
- Ziehn, T., A. Lenton, and R. Law, 2020: An assessment of land-based climate and carbon reversibility in the Australian Community Climate and Earth System Simulator. *Mitigation Adapt. Strategies Global Change*, **25**, 713–731, <https://doi.org/10.1007/s11027-019-09905-1>.