Journal of Hydrology 519 (2014) 3566–3577

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Evaluation of CMIP climate model hydrological output for the Mississippi River Basin using GRACE satellite observations

Frank R. Freedman *, Katherine L. Pitts [1], Alison F.C. Bridger

Department of Meteorology and Climate Science, San Jose State University, One Washington Square, San Jose, CA 95192-0104, USA

ARTICLE INFO

Article history:
Received 22 July 2014
Received in revised form 10 October 2014
Accepted 11 October 2014
Available online 22 October 2014
This manuscript was handled by Konstantine P. Georgakakos, Editor-in-Chief, with the assistance of Ashish Sharma, Associate Editor

Keywords:
Global climate models
Terrestrial water storage
Satellite observations
CMIP5
GRACE
Mississippi River Basin

SUMMARY

We use measurements of terrestrial water storage (TWS) inferred from Gravity Recovery and Climate Experiment (GRACE) satellite observations to evaluate the hydrological output composite-averaged over the Mississippi River Basin (MSRB) and over the ten-year period 2003–2012 from a subset of GCMs from the World Climate Research Programme's Coupled Model Intercomparison Project Phase 5 (CMIP5) and Phase 3 (CMIP3). We then investigate mid-20th to 21st century hydrological trends over the MSRB projected by the CMIP5 models. Improvements were found in CMIP5 simulations of the annual cycle of composite TWS, estimated as the sum of modeled depth-integrated soil moisture and snow water, over those of CMIP3 when compared with the GRACE composite TWS cycle. These improvements coincide with higher horizontal resolution and changes in hydrological parameterizations applied in most of the CMIP5 GCMs compared to earlier CMIP3 versions. Simulated values of composite hydrological budget terms among CMIP5 models, however, are not improved overall, with some models exhibiting increased precipitation and others decreased runoff from CMIP3 to CMIP5 to values outside long-term observed ranges. Since the effect of both increased precipitation and decreased runoff is to increase infiltration and soil water retention, the composite TWS annual cycles from these CMIP5 models, whose earlier CMIP3 simulations in some cases highly underestimated TWS annual cycle amplitudes compared to GRACE, now better agree with GRACE. In spite of the improved prediction of the composite TWS annual cycle, multi-decadal hydrological trends for the MSRB produced by the CMIP5 models vary. A consensus for future decreasing soil moisture is found among models, but with varied responses in magnitude as well as in direction of annual precipitation, evapotranspiration and runoff trends. Overall, GRACE data appear highly useful for evaluating GCM hydrological predictions over large river basins, and a longer time period of these data as more retrievals become available should help to evaluate GCM hydrological output on a multi-decadal time scale.

1. Introduction

Hydrological output fields from global climate models (GCMs) are increasingly being analyzed to determine their suitability to assess future impacts of climate change on water resource availability. Studies have covered both global and regional hydrology, and have investigated most of the earth's major river basins (e.g. Wang, 2005; Milly et al., 2005; Li et al., 2007; Music and Caya, 2007; Schumann et al., 2009; Sperna-Weiland et al., 2012). Data used to evaluate GCM hydrological output fields in these studies include in-situ measurements from well piezometers and river discharge gauges, and remotely sensed data from microwave satellites.

The Mississippi River Basin (MSRB) is the fourth largest river basin in the world. Water from the basin is essential for global food supply, and is the main source for regional use. The river system of the basin is also an important transport conduit for global trade. For these reasons, it is critical to assess how MSRB water resources might be affected by climate change. In addition, because it is such a large basin centrally located in North America, spanning multiple climate regimes, and with good data availability and background understanding from previous scientific work, the MSRB is an excellent testbed for evaluating GCM hydrological fields across large continental areas.

Milly et al. (2005) inferred future changes in total runoff due to climate change within major global river basins by analyzing climate change projections from an ensemble of GCM simulations

* Corresponding author. Tel.: +1 650 387 8926.
E-mail addresses: frank.freedman@sjsu.edu, frank.freedman@sjsu.com (F.R. Freedman), katie.l.pitts@gmail.com (K.L. Pitts), alison.bridger@sjsu.edu (A.F.C. Bridger).
[1] Present address: Department of Atmospheric Sciences, MS3150, Texas A&M University, College Station, TX 77843-3150, USA.

http://dx.doi.org/10.1016/j.jhydrol.2014.10.036
0022-1694/

used in the World Climate Research Programme's Coupled Model Intercomparison Project Phase 3 (CMIP3) multi-model dataset. The climate projections comprising their ensemble came from CMIP3 GCMs whose 20th century control simulations of annual basin-integrated runoff compared best with long-term basin streamflow measurements. Ensemble mean projections showed increased streamflow over the 21st century in Alaska/Northern Canada and decreased streamflow in the desert southwest of North America. Within the MSRB, plots from Milly et al. (2005) showed weak streamflow trends, possibly due to cancellation of trends in sub-regions of the basin. Lu et al. (2010) input 21st century climatological projections of surface temperature and precipitation from an ensemble of CMIP3 GCMs to a regional hydrological model to infer changes in hydrological quantities in the northern MSRB, validating the hydrological model using late 20th century streamflow measurements at the Grafton, Illinois USGS streamflow gauge station. Projected 21st century changes by the regional model were seasonal, which may also explain the weak long-term trends in the MSRB inferred from plots in Milly et al. (2005). The large scatter between modeled runoff and streamflow measurements introduces uncertainties in both studies. Also, streamflow measurements exhibit large local variability, and therefore can misrepresent basin-wide hydrological quantities (Collow et al. 2012).

Since 2002, terrestrial water storage (TWS) anomaly measurements have been made by the Gravity Recovery and Climate Experiment (GRACE) satellite system (Rodell et al., 2007; Famiglietti and Rodell, 2013). TWS measurements from GRACE are spatially complete and depth-integrated. When spatially averaged, GRACE data can therefore represent TWS anomalies over hydrological basins, allowing for more consistent comparison to areally-averaged hydrological output from GCMs than can be achieved using in-situ measurements. Also, since its measurements are depth-integrated, GRACE provides a complete observation of terrestrial water. The data can therefore be used to quantify the storage term in the hydrological balance equation, making the data especially useful for evaluating model hydrological budgets. Microwave satellites, in contrast, only measure soil moisture in the upper several soil centimeters, and inaccuracies in these measurements on the order or 10% volumetric soil moisture have also been reported (Prigent et al., 2005; Collow et al., 2012).

Many previous investigations have used GRACE TWS anomaly data to study regional hydrology and to compare to output from hydrological models. Examples include studies of TWS change in the San Joaquin and Sacramento River basins (Famiglietti et al., 2011), TWS depletion in Central Europe (Anderson et al., 2005), groundwater monitoring in the Central U.S. High Plains aquifer (Rodell and Famiglietti, 2002), and groundwater change in the MSRB basin (Rodell et al., 2007; Zaitchak et al., 2008). Development and evaluation of hydrological models using GRACE data include Niu et al. (2007), Ramillien et al. (2008) and Seneviratne et al. (2010).

In this paper, we compare GRACE observations of TWS anomalies, averaged over the years 2003–2012 and across the MSRB, to identically averaged anomalies of TWS diagnosed from gridded hydrological output fields simulated by a subset of models in the World Climate Research Programme's Coupled Model Intercomparison Project Phase 5 (CMIP5) and earlier Phase 3 (CMIP3) multi-model datasets (Meehl et al. 2007; Taylor et al. 2012). We diagnose model TWS as the sum of depth-integrated soil moisture and snow water, the only two components of TWS in the CMIP archive. A more complete evaluation of one CMIP5 model (CCSM4), also utilizing output fields of explicit groundwater and surface water to diagnose model TWS, was also performed to assess uncertainty due to using only soil moisture and snow water to diagnose model TWS, in light of previous studies demonstrating the importance of various components of TWS in explaining temporal and spatial patterns of major river basins (e.g., Niu et al., 2007; Zeng et al., 2008; Kim et al., 2009; Watanabe et al., 2010; Pokhrel et al., 2012; Pokhrel et al., 2013; Cai et al., 2014). We complement the GCM evaluation by comparing model soil moisture output to spatial patterns of ten-year (2003–2012) average soil moisture across the MSRB produced by the North Americal Regional Reanalysis (NARR) and by comparing ten-year, MSRB-averaged model hydrological budget quantities to observations reported in previous research. By comparing CMIP5 model hydrological cycles of the MSRB to those of earlier CMIP3 model versions, an assessment is made of the improvement that has occurred in recent years in GCM predictions of the MSRB annual hydrological cycle on the basin scale. This will provide a useful baseline for future CMIP-type comparison studies, in which it can be anticipated that more GCMs will contain explicit groundwater compartments in their land surface parameterizations. We finally investigate the multi-decadal trends simulated by CMIP5 models of hydrological variables from the mid-20th to 21st century to study possible future changes in hydrological variables averaged over the MSRB. In addition to its direct purpose of studying possible long-term trends in the basin due to climate change, the analysis will also provide information about the extent to which improvements in GCM simulation of the composite annual hydrological cycle help narrow intermodal variability in GCM simulations of multi-decadal trends of the composite hydrological cycle.

The research in this paper updates previous work of Li et al. (2007), who assessed CMIP3 models against in-situ soil moisture measurements. Here, we investigate both CMIP5 and CMIP3 models. We also utilize for model assessment GRACE TWS measurements, for which a full decade of data is now available to form an observed decadally-averaged composite annual cycle of depth-integrated soil moisture and groundwater anomalies for model comparison.

2. Methodology

2.1. GRACE

The twin satellites GRACE-A and GRACE-B have been in orbit since March 2002. The satellites are in tandem polar orbits, with their speed and relative positions communicated by a K-band microwave ranging system. The microwave measurements are processed into Level 2 data products of the earth's gravitational field in the form of spherical harmonic coefficients at a quasi-monthly temporal resolution. Gravitational perturbations from non-hydrological processes, such as tidal motions and circulations of the atmosphere and ocean, are then removed from GRACE Level 2 data using numerical model output. The resulting fields are then processed into Level 3 gridded fields of TWS anomalies, ice mass changes, ocean bottom pressure changes, and sea level variations. GRACE TWS anomalies thus reflect the total mass change below the surface aggregating all forms of terrestrial water. See Rodell et al. (2007), CSR (2014) and JPL (2013) for more details concerning the GRACE satellites and data processing procedure.

For our analysis, we start with the GRACE Release 5 quasi-monthly, global, 1° latitude × 1° longitude land grids of TWS anomalies (Landerer and Swenson, 2012; Swenson and Wahr, 2006). These gridded data were processed by Sean Swenson, supported by the NASA MEaSUREs Program, and are available at http://grace.jpl.nasa.gov. The data were spatially smoothed using a Gaussian smoother with radius 200 km to remove noisy short wavelength spectral harmonic coefficients. Using a GIS river basin shapefile, we extracted from these global grids the values within the MSRB for the ten-year period January 2003–December 2012.

For each point within the MSRB, the quasi-monthly GRACE TWS time series was then piecewise linearly interpolated to the regular monthly time series of the CMIP model output. The average of the resulting GRACE time series at each spatial point within the MSRB was then subtracted from each value in the time series to obtain anomalies relative to a 2003–2012 base. The monthly anomaly time series was then spatially averaged, and then averaged in time for each calendar month over the ten year period to form a ten-year, MSRB-averaged composite TWS anomaly annual cycle. This composite was compared to CMIP model TWS outputs processed in an identical manner, as described below. Applying the procedure given in JPL (2013), as described in more detail in Section 2.2, we calculated quasi-monthly GRACE TWS anomalies averaged over the MSRB to be accurate to approximately 11 mm liquid water, or equivalently a 1.1% accuracy of volumetric soil moisture over a one-meter soil depth.

The monthly, MSRB-averaged composite GRACE TWS time series can be explained by the following equation,

$$(\mathrm{TWS})_{n+1} - (\mathrm{TWS})_n = \mathrm{P}_n - \mathrm{E}_n - \mathrm{R}_n, \quad (1)$$

where (TWS) is the composite TWS anomaly; P, E, and R are the monthly precipitation, evapotranspiration and total (surface plus drainage) runoff, respectively, averaged over the area of the basin; and subscripts (n) and ($n+1$) represent a given and next month, respectively. The units of the terms in (1) are millimeters. TWS comprises terrestrial water in all of its components – biomass, surface water, ice, snow, soil moisture and groundwater – however GRACE by itself is not able to reconcile these individual components. To do this, researchers (e.g. Rodell et al., 2007; Kim et al., 2009) have utilized offline land surface modeling, comparing TWS from the model to GRACE data in order to validate model TWS. While Rodell et al. (2007) found that in the MSRB changes in soil moisture and groundwater mainly determine GRACE TWS changes, Kim et al. (2009) found important contributions to MSRB basin TWS anomalies from surface water. As described below, our estimate of TWS from CMIP models only includes soil moisture and snow water since these are the only components of TWS provided in the CMIP archive, however for one model we will compare results utilizing explicit groundwater and surface water fields to check the sensitivity of our analysis to this assumption.

2.2. CMIP models

The GCMs chosen for analysis are listed in Table 1. Shown in the table for each model are the number of runs comprising their ensemble, the horizontal grid resolution, the depth and number of layers in the model's soil column, a summary of the main changes to the model's hydrological parameterization in the CMIP5 compared to CMIP3 version, and a literature reference. We chose this subset of models because they have multi soil-layer land-surface parameterizations and because output contributions are available for both CMIP5 and CMIP3, allowing the performance of the two model versions to be compared. The hydrological parameterizations for these models all have advanced process-based formulations for evapotranspiration and runoff, with monthly output of these quantities archived. Along with archived monthly surface precipitation rates, the terms on the right-hand side of (1) for the models can then be evaluated. As seen from Table 1, most of the CMIP5 models have increased their horizontal resolutions and have multiple runs comprising their ensembles in comparison to the CMIP3 versions. Hydrological parameterizations have also changed for several models, either or both by switching to a different scheme or by increasing the depth or number of layers in their soil column. All models except HadCM3 and the INM GCMs have river routing. Explicit representations of groundwater are included in CCSM4 and GFDL-CM3, with remaining models representing groundwater implicitly as the saturated portion of their soil column.

We estimate TWS from CMIP models as the sum of soil moisture and snow water, since these are the only components of TWS provided in the CMIP archive. Monthly outputs for the chosen GCMs were obtained from the CMIP5 and CMIP3 data archives. We used the variable 'mrso' for soil moisture, which is the model output monthly averaged soil moisture in all phases summed over all model soil layers and averaged over the land portion of the grid cell. We used the variable 'snw' for snow water, also the sum over all snow layers and grid area averaged. The units of this variable are millimeters. We used output fields of monthly average precipitation rates in all phases ('pr'), monthly average surface latent heat flux ('hfls'), and monthly average total runoff ('mrro') for terms on the right-hand side of (1). After dividing 'hfls' by the latent heat of vaporization, the units of these terms are all mm/s, which we convert to monthly accumulations (in millimeters) assuming 30.4 (= 365/12) days per month. We processed the model output fields for these variables in the same manner as used for the GRACE data – by extracting the portion of the gridded field within the MSRB, expressing them as anomalies relative to the 2003–2012 average, averaging the time series over the MSRB, and then averaging the resulting MSRB-averaged monthly time series for each calendar month to form a ten-year, MSRB-averaged composite annual cycle. We then averaged the composites over all model runs (when there are more than one) to form ensemble average composites.

For CMIP5 models, we constructed the 2003–2012 time series by connecting output fields from the 'historical' 20th century cli-

Table 1
CMIP5 and CMIP3 GCMs analyzed and associated model information. CMIP5 models and information are on left, and CMIP3 models and information are on right in parenthesis. Information obtained from the IPCC AR5 Working Group 1 Report (IPCC, 2013), the "Model Documentation" section of PCMDI (2010), information provided in the source listed under 'Reference', and papers referenced therein.

Model	Runs	Grid resolution		Soil column		Hydrological Parameterization	Reference
		Lat (°)	Long (°)	Layers	Depth (m)	Major changes since CMIP3	
HadGEM2-ES (HadGEM1)	4 (1)	1.25 (1.25)	1.88 (1.88)	4 (4)	3 (3)	Various improvements	Martin et al. (2011)
IPSL-CM5A-LR (-CM4)	4 (1)	1.875 (2.5)	3.75 (3.75)	2 (2)	4 (2)	Deeper soil layer	Dufresne et al. (2013)
MIROC5 (-3.2-medres)	3 (3)	1.4 (2.8)	1.4 (2.8)	6 (5)	14 (4)	Deeper soil layer	Watanabe et al. (2010)
HadCM3	10 (1)	2.5	3.75	4	3	No changes[b]	Cox et al. (1999)
MRI-CGCM3 (-CGCM2.3.2)	1 (5)	1.125 (2.8)	1.125 (2.8)	9/11[a] (3)	1.5/3.5[a] (1.5/3.5)	Model change: HAL	Yukimoto et al. (2011)
GISS-E2-R (-ER)	3 (1)	2 (4)	2.5 (5)	6 (6)	3.5 (3.5)	No changes	Schmidt et al. (2006)
GFDL-CM3 (-CM2.0)	1 (1)	2 (2)	2.5 (2.5)	20 (18)	10 (6)	Deeper soil layer model change: LM3[c]	Milly et al. (2014)
CCSM4 (CCSM3)	6 (5)	0.9 (1.4)	1.25 (1.4)	10 (10)	3.8 (3.4)	Model change: CLM4[c]	Lawrence et al. (2011)
INM-CM4 (-CM3.0)	1 (1)	1.5 (4)	2.0 (5)	23 (23)	10 (10)	No changes[b]	Volodin et al. (2010)

[a] Number of soil layers and depth of soil column in MRI models vary with surface vegetation type. Numbers on left and right of '/' refer to values used for grassland and forested grids, respectively, the predominant types covering the MSRB.
[b] HadCM3 and INM GCMs do not contain a river routing scheme.
[c] CCSM4 and GFDL-CM3 hydrological parameterizations include an explicit groundwater compartment.

mate reconstruction, which ends at 2005, and 'rcp85' future climate projection, which begins at 2006. An exception is the CMIP5 HadCM3 model, for which we used the 'rcp45' scenario, the only 'rcp' scenario archived for this model. For CMIP3 models, we use the A2 21st century climate projection, which begins at 2000. The ten-year (2003–2012) composite MSRB-averaged TWS cycles among the various climate forcing scenario runs for a given model do not differ greatly based on our inspection of plots of the archived model output (not shown). Since the A2 and 'rcp85' emission scenarios assume continuation of current-day greenhouse gas emission trends, we chose these for presentation in this paper so that our analysis of model long-term hydrological trends (Section 3.5) assesses a worst-case, "business as usual" situation.

3. Results

3.1. CMIP models versus GRACE: results

GRACE TWS anomalies averaged over the MSRB are shown in Fig. 1. The plot shows the ten-year series of monthly values, created by piecewise linearly interpolating the original quasi-monthly data. While relatively wet and dry years can be seen, notably the recent drought year of 2012, little trend over the decade is apparent. The inset shows the GRACE ten-year composite annual cycle (black line). The composite reveals a maximum in April of around +50 mm, and a minimum in September of around −60 mm. Interannual peak values, however, can be different by as much as approximately a factor of two for a particular year. This is due both to phase and amplitude differences in the annual cycles from year to year. Also plotted in Fig. 1 are the CMIP5 (blue dashed) and CMIP3 (red dashed) multi-model mean monthly time series and composite annual TWS cycles simulated by the chosen GCMs in Table 1. The composite multi-model means agree well with GRACE, with improvement from CMIP3 to CMIP5. Note that the monthly model time series do not have much interannual variability, since these time series are the result of averaging over results from several models and multiple runs comprising individual model ensembles.

The amount of TWS that on average builds up during fall and winter due to excess precipitation and is later released to the atmosphere and oceans during spring and summer via evapotranspiration and runoff is indicated by the difference of TWS from peak to trough in the annual composite curve. Defining this quantity as ΔTWS, we therefore evaluate it as,

$$\Delta\mathrm{TWS} = \max(\overline{\mathrm{TWS}}) - \min(\overline{\mathrm{TWS}}), \tag{2}$$

where the overbar represents the composite average, and 'max' and 'min' are the maximum and minimum monthly values, respectively, comprising the composite annual cycle. Evaluating (2) using the GRACE TWS composite yields $\Delta\mathrm{TWS} \approx 120$ mm. Analyses of measurements from the last half of the 20th century (Milly, 2005; Qian et al., 2007; Music and Caya, 2007; Cai et al., 2014) find values for MSRB average annual precipitation of approximately 800 mm and average annual runoff/discharge of approximately 200 mm. The average annual amount of water cycling into and out of the soil over the MSRB is thus approximately 15% of the basin annual average precipitation ($\Delta\mathrm{TWS}/\mathrm{P} \approx 0.15$) and approximately 60% of the basin annual average runoff/discharge ($\Delta\mathrm{TWS}/\mathrm{R} \approx 0.6$).

The composite annual cycles of CMIP5 and CMIP3 individual model TWS anomalies along with the corresponding GRACE composite are shown in Fig. 2. Multi-model mean composite cycles of both the CMIP5 and CMIP3 models reproduce well the seasonal variation of the GRACE TWS anomaly, with the CMIP5 multi-model mean better matching GRACE. Intermodel variability is noticeable, however, especially with respect to amplitudes. The variability is less for CMIP5 compared to CMIP3 models, and correspondingly improvements versus GRACE are found for most of the individual CMIP5 model simulations compared to those of CMIP3. Major performance improvements of the CCSM, INM and GISS models especially contribute to the improved performance of the CMIP5 multi-model mean. Other models that show significant improvement are GFDL and IPSL.

The model versus GRACE mean absolute errors (MAE) and correlation coefficients (r) are given for each model in Table 2. These are computed using the twelve monthly values of the model and GRACE TWS composites as follows,

$$\mathrm{MAE} = \frac{1}{12}\sum_{n=1}^{12}\left|(\overline{\mathrm{TWS}})_n^{\mathrm{mod}} - (\overline{\mathrm{TWS}})_n^{\mathrm{G}}\right|, \tag{3}$$

$$r = \sum_{n=1}^{12}(\overline{\mathrm{TWS}})_n^{\mathrm{mod}}(\overline{\mathrm{TWS}})_n^{\mathrm{G}} \Big/ (s_{\mathrm{mod}}s_{\mathrm{G}}), \tag{4}$$

where superscripts 'mod' and 'G' indicate model and GRACE quantities, respectively, and 's' represents the standard deviation of the twelve monthly composite values. We also give in Table 2 corresponding values of 'amplitude ratio', defined as the absolute area under each model composite curve divided by that under the GRACE composite curve, to distinguish model overprediction (value > 1) and underprediction (value < 1) of amplitude. MAE values decrease from CMIP3 to CMIP5 for all GCMs except three (HadGEM2, MRI and MIROC). The CMIP5 models with especially big improvement are CCSM4, INM-CM4 and GISS-E2-R, with CCSM4 now giving a very low MAE (3.5 mm) in contrast to the poor score corresponding to its earlier CMIP3 version (29.4 mm). CCSM4 produces the lowest MAE among CMIP5 models, and IPSL-CM5A-LR produces the highest (19.7 mm). Correlation coefficients are high for both the CMIP3 and CMIP5, generally greater than 0.9, with improvement evident with CMIP5. The high 'r' values are a consequence of the dominance of the annual cycle of TWS, which all models reproduce well. Amplitude ratios also improve from CMIP3 to CMIP5, with CMIP5 values now closer to the correct value of one compared to CMIP3. Six of the CMIP5 models underpredict amplitudes (ratio < 1) and three overpredict amplitudes (ratio > 1). The

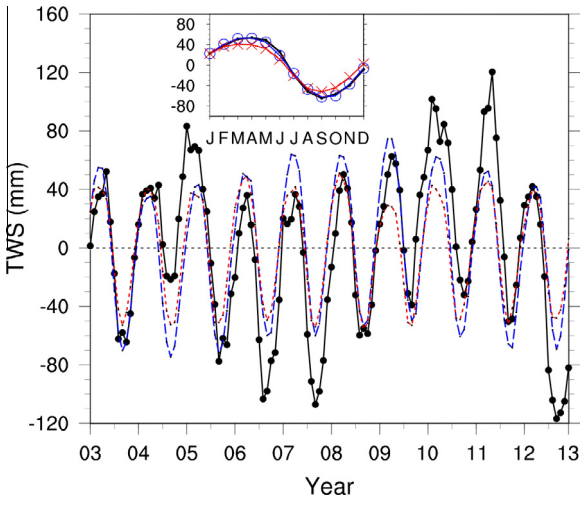

Fig. 1. Monthly GRACE TWS anomalies averaged over the MSRB from 2003 to 2012 (black solid) versus CMIP5 (blue long dashed) and CMIP3 (red short dashed) multi-model mean anomalies simulated by the GCMs in Table 1. The inset shows the corresponding composite TWS annual cycles (markers added to model lines for clarity).

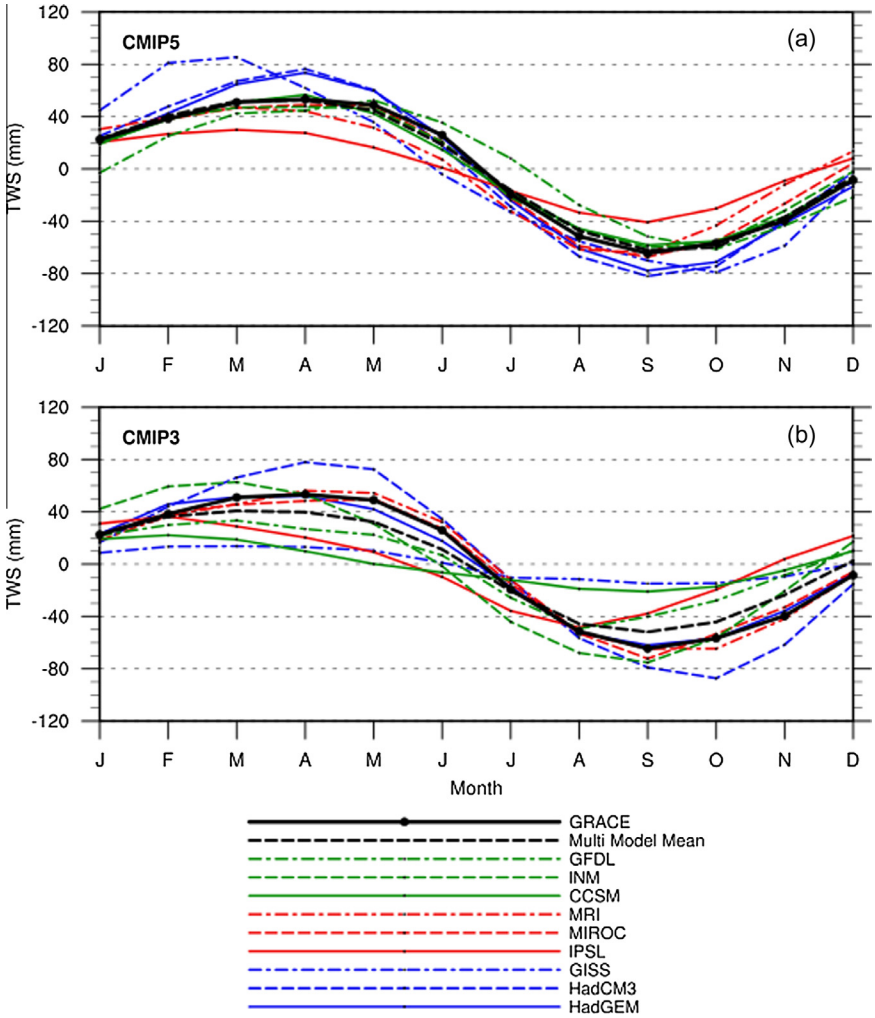

Fig. 2. Annual cycles of (a) CMIP5 and (b) CMIP3 model simulations of composite TWS anomalies averaged over the MSRB. Composite is over the period 2003–2012. The multi-model mean cycles are shown by the dashed black line. The GRACE TWS average composite is shown by the solid, black line.

Table 2
Mean absolute error (MAE), correlation coefficient (r), and amplitude ratio (see Section 3.1 for definition) of simulated CMIP5 and CMIP3 model TWS anomaly composite annual cycles versus GRACE TWS anomaly composite cycles averaged over the MSRB. Composites are over the ten-year period 2003–2012. CMIP5 values on left, CMIP3 on right in parenthesis. Multi-model means shown on bottom row (bold italicized).

Model	MAE[a] (mm)	*r*	Amplitude ratio
CCSM4 (CCSM3)	3.5 (29.4)	0.995 (0.798)	0.93 (0.33)
INM-CM4 (CM3.0)	4.1 (16.2)	0.997 (0.929)	0.91 (1.11)
MIROC5 (3.2-medres)	4.6 (3.3)	0.990 (0.996)	0.96 (0.97)
HadGEM2-ES (HadGEM1)	8.5 (3.0)	0.997 (0.996)	1.20 (0.97)
HadCM3	11.4 (14.2)	0.992 (0.988)	1.23 (1.30)
MRI-CGCM3 (CGCM2.3.2)	12.1 (3.7)	0.943 (0.995)	0.89 (1.02)
GFDL-CM3 (CM2.0)	12.9 (17.5)	0.937 (0.927)	0.87 (0.63)
GISS-E2-R (ER)	18.0 (30.0)	0.945 (0.966)	1.29 (0.25)
IPSL-CM5A-LR (CM4)	19.7 (24.8)	0.942 (0.763)	0.54 (0.63)
Multi-Model Mean	***2.8 (9.7)***	***1.000 (0.986)***	***0.96 (0.77)***

[a] Uncertainty in MAE values estimated to be within 3 mm. See Section 3.2 for details.

predominance of model under-rather than overprediction of amplitude is also seen in the CMIP3 results.

Overall, the improvements from CMIP3 to CMIP5 appear associated with both increased horizontal resolution and changes to model hydrological parameterizations. Better horizontal resolution aligns positively with improved model performance for CCSM4, INM-CM4, GISS-E2-R, and IPSL-CM5A-LR. The best indication of this is with INM-CM4 and GISS-E2-R, models for which horizontal resolution was especially coarse in CMIP3 runs and for which there were no changes in hydrological schemes from CMIP3 to CMIP5. Concerning scheme changes, the best example of the positive effects is CCSM4, where significant improvements were incorporated into Community Land Model (CLM4). GRACE measurements, in fact, were utilized in model evaluation of the interim CLM version 3.5 (Oleson et al., 2008). Use of deeper soil layers in CMIP5 versions are associated with improved MAE scores in IPSL-CM5A-LR and GFDL-CM3, indicating a role in this factor as well. Some counterexamples to these associations exist, for example the worse performance of the HadGEM2-ES and MRI-CGCM3 compared to earlier CMIP3 versions in spite of changes to hydrological schemes and increased horizontal resolution (in the case of MRI-CGCM3), and the little change of MIROC5 results in spite of the increase in horizontal resolution and soil depth. These counterexamples, however, are all cases where CMIP3 models were performing relatively well already, and any degraded performance in MAE scores of CMIP5 versions was modest.

3.2. CMIP models versus GRACE: uncertainty analysis

We focus our uncertainty analysis on MAE. Uncertainties of MAEs in Table 2 were checked with respect to errors in GRACE data and uncertainties in comparing the observed GRACE composite to

model composites generated from a single model run. We also examine the error due to using only soil moisture and snow water to diagnose model TWS by comparing simulations from CCSM4 (obtained outside the CMIP archive), for which explicit groundwater and surface water fields are available.

GRACE data contain measurement and leakage error, which are additive for a spatial data point. Over an area, however, the error calculation must account for spatial correlations of pointwise errors. Applying the procedure given in JPL (2013), we calculated a total (measurement plus leakage) error averaged over the MSRB of 11.0 mm, which applies to the individual quasi-monthly points in the GRACE MSRB-averaged time series. Interpreting this value as twice the standard error of an individual measurement in the series, we then created 100 alternative GRACE MSRB-averaged monthly time series by using a random number generator to construct a series of 12,000 (120 monthly points times 100 series) perturbations, with mean zero and twice the standard deviation of 11 mm. These were then added in sequence (120 points at a time) to the piecewise linearly interpolated monthly points of the measured GRACE MSRB-averaged time series. The average of each of these 100 new time series was then subtracted from the original series to obtain time series with average anomaly of zero. The composites for each of the 100 series were then constructed, and the MAEs of each CMIP5 TWS composite relative to each of these 100 series were then calculated. This procedure yields 100 MAEs for each CMIP5 model, from which the mean and standard deviation of MAEs was then calculated for each model. Twice the average standard deviation of the MAEs across the nine models was found to be approximately 1.0 mm, which we take as the uncertainty in model MAEs relative to GRACE due to GRACE measurement error.

Only one 'rcp85' CMIP5 run was available for INM-CM4, MRI-CGCM3 and GFDL-CM3. In CMIP3, only one run was available for all models except CCSM3, MIROC3.2-medres and MRI-CGCM2.3.2. To quantify the uncertainty in MAEs when one run, rather than the ensemble mean, is used to compare to GRACE, we calculated the variability of the MAEs for the ten runs of the CMIP5 HadCM3 model, and for the six runs of the CCSM4 model. Twice the standard deviation of the MAEs from the ten HadCM3 runs was 5.2 mm, and the difference between the maximum and minimum MAEs of these runs was 8.0 mm. Twice the standard deviation of the MAEs from the six CCSM4 runs was 1.7 mm, and the difference between the maximum and minimum MAEs of these runs was 2.0 mm. An uncertainty up to around 5 mm when one model run is compared to GRACE can thus roughly be inferred. Because of the smoothing effect of averaging, the uncertainty would be smaller than this as the number of runs to form an ensemble mean increases.

To investigate the error due to the use of only soil moisture and snow water to estimate model TWS, we obtained for the CCSM4 model (from http://www.earthsystemgrid.org outside the CMIP archive) gridded fields of explicit groundwater and surface water for the model's CMIP5 rcp6.0 future scenario, the only model run for which these fields were available. The groundwater fields are the CLM4 variable 'WA', defined as water stored in the unconfined aquifer (mm), and the surface water grids are the CLM4 variable 'VOLR', defined as the storage of river water (m^3, which we converted to mm by dividing by the grid box area) associated with the model's river routing scheme (Oleson et al., 2010).

The model composite (2006–2012, MSRB-averaged) individual components, sum of soil moisture and snow water, and total over all components along with the GRACE TWS (2003–2012) are shown in Fig. 3. The MAE versus GRACE of the soil moisture plus snow composite is 4.5 mm, and the MAE for the total (over all components) composite versus GRACE is 8.5 mm. This suggests an error of around 4 mm (the difference between the two) from using

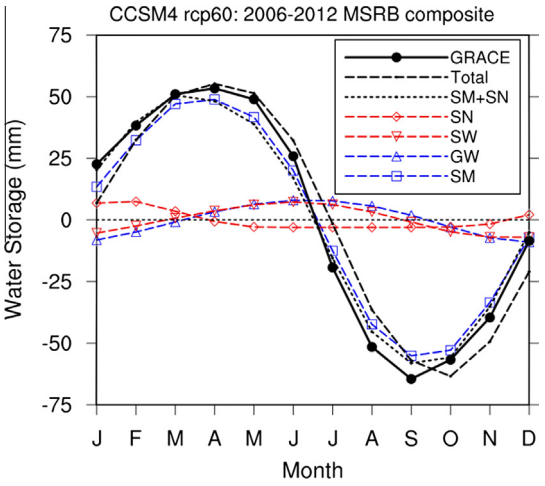

Fig. 3. Annual cycles of composite TWS and its components for the CCSM4 model rcp6.0 runs. Composites are over the period 2006–2012. Components are depth-integrated soil moisture (SM), snow water amount (SN), explicit groundwater (GW) and surface water (SW). Also shown are composites of soil moisture plus snow water (SM + SW) and the total of all components ('Total'). The GRACE TWS composite over the period 2003–2012 is also shown.

only the soil moisture and snow components in our analysis. Of course, this estimate is uncertain since it is deduced from one set of runs for one particular model. In fact, the surface water composite over the MSRB shown in Kim et al. (2009), using the MATSIRO land surface model (used in MIROC5), is quite different than shown in Fig. 3 for CCSM4, with an amplitude in April of around 20 mm, about twice as high and earlier in the spring than shown in Fig. 3. While the difference in MAE versus GRACE using soil moisture plus snow as opposed to using all components is thus very likely model dependent, we still expect the error to be modest, around 4 mm or slightly higher. Indications (Rodell et al., 2007; Kim et al., 2009) are that soil moisture and groundwater anomalies comprise the majority of TWS anomalies in the MSRB, with most of the groundwater anomaly implicitly accounted for in modeled soil moisture as long as model soil columns are deep enough for the water table to not fall beneath the bottom model soil level.

Taken collectively, these findings suggest that differences within 5 mm in MAEs versus GRACE among models, and between CMIP3 and CMIP5 versions of the same model, are likely insignificant. This value, suggested from the analysis above, is also corroborated by the difference of around 3 mm between CMIP3 and CMIP5 runs of HadCM3 (Table 2), a model that had not changed from CMIP3 to CMIP5. As a consequence, the improvement in the multi-model mean MAE from 9.7 mm in CMIP3 to 2.8 mm in CMIP5 is likely significant. This is associated with likely significant reductions in MAEs from CMIP3 to CMIP5 in four of the nine models: CCSM4, INM-CM4, GISS-E2-R and IPSL-CM5A-LR. Three of the models (INM-CM4, CCSM4 and MIROC5) have MAEs that are likely significantly lower than the CMIP3 multi-model mean of 9.7 mm, four (HadCM3, MRI-CGCM3, GFDL-CM3, and HadGEM2-ES) have MAEs not significantly different, and two (GISS-E2-R and IPSL-CM5A-LR) have MAEs that are likely significantly higher. Comparing the CMIP5 models among themselves, the difference of MAEs between the two models with lowest MAEs (INM-CM4 and CCSM4) and highest MAEs (IPSL-CM5A-LR and GISS-E2-R) is likely significant, while the differences in MAEs between HadGEM2-ES, HadCM3, MRI-CGCM3, and GFDL-CM3 are likely insignificant.

Finally, while our decision to use spliced CMIP5 'historical' and 'rcp85' runs in our analysis was made so that long-term model 21st century trends could also be investigated, it is worth briefly looking into whether improved MAE scores could be achieved instead

using the CMIP5 'decadal' runs. These are initialized closer in time to the GRACE data period, and therefore may give better predictions versus GRACE. We used the archived 'decadal' runs (six ensemble members) for MIROC5 initialized in 2002, a year before the time period of the GRACE observations. We did not apply a drift correction (CLIVAR, 2011) to the monthly model TWS (soil moisture plus snow) output from the 2002 'decadal' run before processing it into a ten-year MSRB-averaged composite. This was because very little drift over ten years is evident in the MIROC5 'decadal' MSRB-averaged TWS time series, as deduced by us by inspecting the plot (not shown) of monthly MSRB-averaged TWS over ten years averaged over eleven MIROC5 'decadal' runs initialized from 1992 to 2002. The MAE versus GRACE for the TWS composite (ensemble-averaged) corresponding to the MIROC5 2002 'decadal' run was 5.6 mm, similar to the value of 4.6 mm for MIROC5 using spliced 'historical' and 'rcp85' runs (Table 2). While, this brief exercise suggests that 'decadal' runs offer little predictive improvement in terms of the 2003–2012 ensemble TWS composite over the MSRB (the focus of this paper), a more thorough investigation looking at multiple models and variation among ensemble members would help answer this question of improved model performance of 'decadal' runs more definitively.

3.3. CMIP5 models: spatial soil moisture patterns

To further evaluate model simulations, we examine CMIP5 model spatial patterns of depth-averaged soil moisture ('mrso' in the CMIP archive, subsequently abbreviated SM) across the MSRB. Contour maps for each model's ten-year (2003–2012) average SM anomaly relative to its ten-year MSRB average value are shown in Fig. 4. Contouring anomalies rather than the full depth-integrated values facilitates intermodel comparison since, because of the different model soil depths (see Table 1), large differences of full values among models exist. We compare the model fields to the corresponding ten-year average field produced by the National Center for Environmental Prediction North America Regional Reanalysis (NARR). The NARR (Mesinger et al., 2006) output, starting from 1979, is a monthly spatial reproduction of two-meter depth integrated SM across North America generated by assimilating observations of surface wind and precipitation every three hours into runs of the NCEP Eta model coupled to the Noah land-surface model. While the NARR therefore does not directly incorporate soil moisture observations, the incorporation of standard surface meteorological and other observations provides observational constraints in the NARR SM field not present in the GCM runs. Also, the Eta model used in the NARR operates on a 32 km horizontal resolution, higher than that of any of the CMIP GCMs (Table 1). The NARR is therefore likely a more accurate framework than the GCMs for representing MSRB SM spatial patterns. Miguez-Macho et al. (2008) compared the long-term average NARR SM field over North America to the SM reanalyses produced by two hydrological models with higher horizontal resolution than the NARR and explicit groundwater compartments. The NARR field was found to be generally similar to those produced by these two more detailed reanalyses.

The NARR field (Fig. 4, upper-left panel) exhibits a general west to east increase in SM across the MSRB of about 200–300 mm. Most models reproduce this general pattern. The CMIP5 models that reproduce this pattern best are CCSM4, HadGEM2-ES and, although with less spatial resolution, HadCM3. The GFDL-CM3 model produces a west to east increasing field with too high a gradient, of approximately 800 mm. The INM-CM4 and MIROC5 models produce overly wet soils in the northwest portion of the basin. The MRI-CGCM3 model is too dry in the eastern part of the basin. The IPSL-CM5A-LR model is too wet in the western part of the basin and lacks spatial detail. Curiously, the GISS-E2-R model has a quite different pattern than the other models and NARR, being most wet in the central part of the basin and dryer to the west and east. This pattern was also seen (not shown) with the CMIP3 version of the GISS model.

The spatial ten-year average precipitation patterns of the models (not shown) appear similar among themselves and to the NARR precipitation pattern, with gradually increasing precipitation across the basin from west to east. Since the general west-to-east SM increase across the basin seen among most of the models in Fig. 4 is, as expected, consistent with this, more subtle differences in the model SM spatial patterns are likely caused by other factors. Model differences in representing evapotranspiration, runoff and infiltration are likely involved, associated with model differences in hydrological parameterization, soil depth and horizontal resolution discussed in Section 3.1. Differences in model ability to accurately represent distributions of precipitation on seasonal and daily time scales are also likely important, since runoff and infiltration rates are affected by precipitation intensity.

3.4. CMIP models: analysis of hydrological budget

In Table 3 we show for each model as well as the CMIP5 and CMIP3 multi-model means the computed composite annual average monthly precipitation (P), evapotranspiration (E), total runoff (R), difference from peak to trough in model composite TWS cycle (ΔTWS), and ratios R/E and ΔSM/R. Observed ranges for P, E, and R based on observational analysis in previous literature cited in the table are given on the bottom row of the table. The observed value $\Delta TWS \approx 120$ mm was discussed in Section 3.1. The tabulated observed value $R/E \approx 0.3$ is based on the midpoint of the observed ranges of R and E, and the observed value $\Delta TWS/R \approx 0.6$ follows from GRACE observations of ΔTWS and the midpoint of the observed range of R.

There is an increase from CMIP3 to CMIP5 of over 100 mm in average annual precipitation for the GFDL-CM3, GISS-E2-R, MIROC5 and MRI-CGCM3 models, with values for GFDL-CM3, GISS-E2-R and MRI-CGCM3 now over 1000 mm. As a result, the multi-model mean average annual precipitation increases from around 860 mm in CMIP3 to around 930 mm in CMIP5. Observed values (bottom row of table), however, are in the range 750–850 mm, and from inspection of Fig. 2 of Milly (2005) in only a few years over the last half century has the average basin annual precipitation approached or exceeded 1000 mm. It thus appears that a high precipitation bias over the MSRB has developed in these CMIP5 models. There are corresponding increases in the simulated composite E and R values by these models. For E, this contributes to an overpredicted multi-model mean evapotranspiration (730 mm) in the CMIP5 models compared to the observed range (550–650 mm).

For R, on the other hand, the multi-model mean value ($\approx$190 mm) of the CMIP5 models remains in the observed range (180–240 mm). This is because other models (INM-CM4, IPSL-CM5A-LR and CCSM4) exhibit moderate to strong decreases in runoff compared to their earlier CMIP3 simulations. In only one of these three models (IPSL-CM5A-LR), however, is the CMIP5 R value within the observed range, with the other two model values lower than observed. Yet, it is interesting that these three models of decreased R are ones with significant improvements in their composite annual TWS cycle compared to GRACE data. The decreased runoff by these models therefore seems to improve the annual cycles of infiltration and soil water retention simulated by these models, leading to better composite TWS anomaly predictions. The decreased R values causes less spread in model simulated values of ratio R/E from CMIP3 to CMIP5. While in both cases the multi-model mean ratio is close to the observed value of around

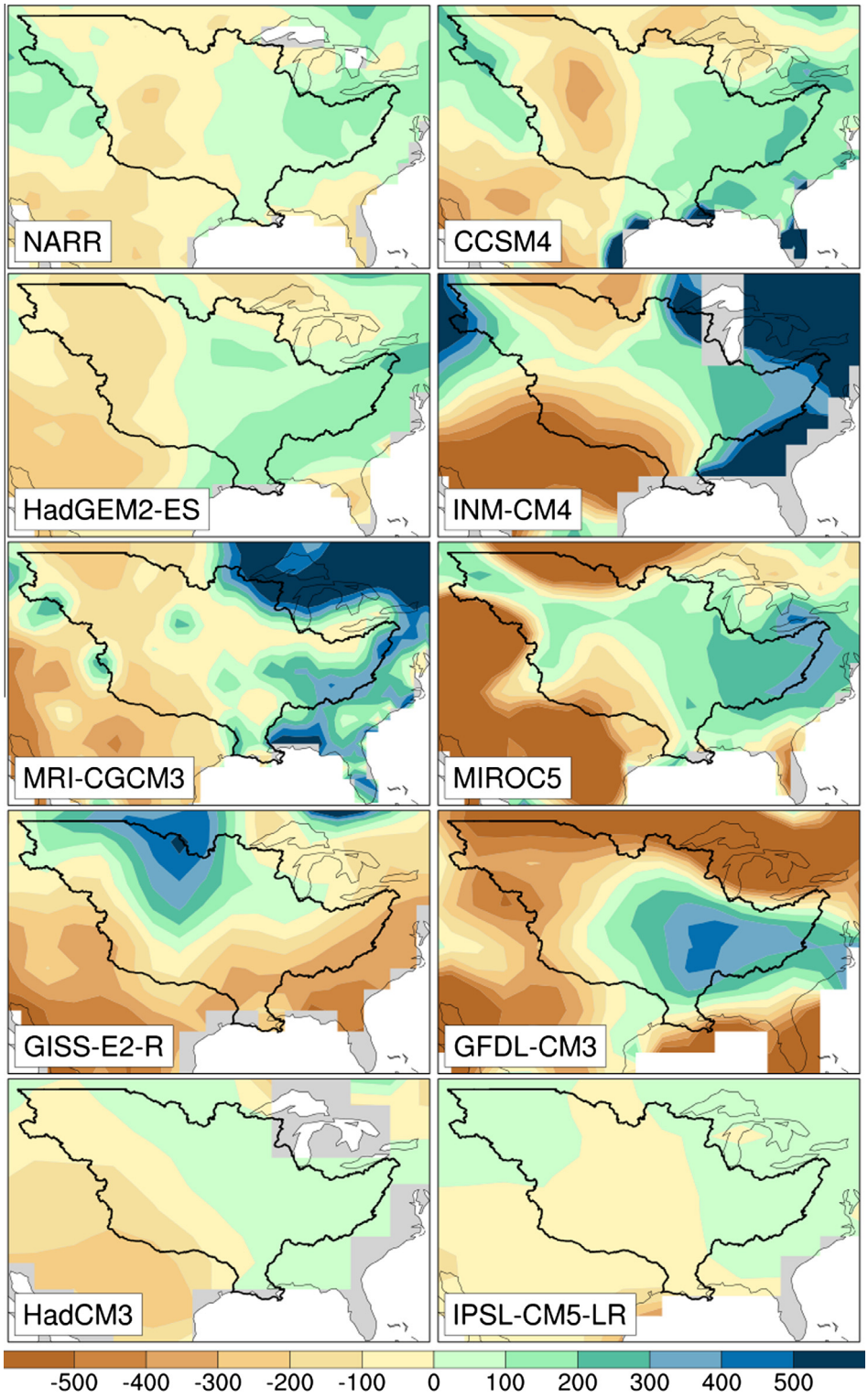

Fig. 4. Spatial distribution across the MSRB of annual SM anomalies (mm) from the North American Regional Reanalysis (NARR) and CMIP5 model simulations. MSRB outlined in black. Anomalies are relative to each model's respective spatial mean over the MSRB over the period 2003–2012.

0.3, the model values range from 0.1 to 0.4 in CMIP5, compared to the range 0.1–0.7 found in CMIP3.

Looking at model simulated values of ΔTWS, it can be deduced that the increased P in some models and decreased R in others, mentioned above, both appear to lead to better simulation of corresponding SM composites of these CMIP5 models. The increased P in GFDL-CM3 and GISS-E2-R increases the amount of water available for infiltration, increasing the model ΔTWS towards better agreement with the observed GRACE value. In association, the ratio ΔTWS/R for these models is in better agreement with the observed value of around 0.6 based on GRACE. The decreased R (with P roughly unchanged) simulated by CCSM4, INM-CM4 and IPSL-CM5A-LR similarly enables more water to infiltrate and/or be retained in the soil, increasing the values of ΔSM simulated

Table 3
Simulated values of annual depth-integrated soil moisture (SM), precipitation (P), evapotranspiration (E), total runoff (R), R/E and ΔSM/R by each CMIP5 and CMIP3 model, averaged over MSRB and GRACE observational period (2003–2012). CMIP5 values on left, CMIP3 on right in parenthesis. Multi-model means and observations on bottom two rows (bold italicized).

Model	P (mm)	E (mm)	R (mm)	ΔTWS (mm)	R/E	ΔTWS/R
CCSM4 (CCSM3)	812 (767)	721 (518)	97 (134)	115 (43)	0.13 (0.26)	1.19 (0.32)
INM-CM4 (CM3.0)	929 (902)	790 (594)	149 (331)	107 (138)	0.19 (0.56)	0.72 (0.42)
MIROC5 (3.2-medres)	914 (788)	748 (721)	166 (79)	116 (122)	0.22 (0.11)	0.70 (1.54)
HadGEM2-ES (HadGEM1)	831 (815)	664 (674)	167[a] (157)	151 (114)	0.25[a] (0.23)	0.90 (0.73)
HadCM3	965 (930)	689 (678)	280 (258)	158 (165)	0.41 (0.38)	0.56 (0.64)
MRI-CGCM3 (2.3.2)	1044 (779)	746 (624)	305 (151)	111 (121)	0.41 (0.24)	0.36 (0.80)
GFDL-CM3 (CM2.0)	1068 (958)	842 (763)	218 (208)	114 (83)	0.26 (0.27)	0.52 (0.40)
GISS-E2-R (ER)	1051 (925)	762 (623)	289 (241)	165 (29)	0.38 (0.39)	0.57 (0.12)
IPSL-CM5A-LR (CM4)	781 (868)	613 (480)	190 (331)	71 (85)	0.31 (0.69)	0.37 (0.26)
Multi-model mean	***933 (859)***	***731 (626)***	***193 (211)***	***115 (93)***	***0.26 (0.33)***	***0.60 (0.44)***
Observations[b]	***750–850***	***550–650***	***180–240***	***≈120***	***~0.3***	***~0.6***

[a] For HadGEM2-ES, runoff is estimated as P–E since variable 'mrro' is not archived.
[b] Reported ranges of P, E, R based collectively on Milly (2005), Music and Caya (2007), Qian et al. (2007) and Cai et al. (2014). Observed value of ΔTWS based on GRACE data. Values of R/E and ΔSM/R based on midpoint of reported ranges of E and R and the GRACE measured value of ΔTWS.

by these models compared to their CMIP3 versions. The ratio ΔTWS/R for these models correspondingly increases compared to CMIP3, closer to the observed value of around 0.6 for the INM-CM4 and IPSL-CM5A-LR models, yet to a value of around 1.2 for CCSM4, substantially greater than the observed value. This is likely due to underpredicted R by the model, discussed above.

The better simulation in CMIP5 TWS composites therefore appears to be strongly influenced by different changes among models from CMIP3 to CMIP5 in the values of hydrological forcing terms, which are not always themselves in better agreement with observations. Alongside the changes in hydrological forcing are different model responses due to the changes in hydrological parameterizations, increased soil depth and increased horizontal resolution, as discussed in Section 3.1. The effect of this on the CMIP5 multi-model mean values of hydrological forcing terms is correspondingly mixed, with multi-model mean values of R, ΔTWS, R/E and ΔTWS/R in better agreement, and multi-model mean values of P and E in worse agreement, respectively, with observed values compared to CMIP3 simulated values.

This type of model evaluation is possible due to the availability for the first time of enough years of GRACE TWS measurements to form a representative TWS composite for the basin. The deduction from this composite of an observed value for the ratio ΔTWS/R is especially useful for model evaluation, and can be used as well to calibrate the relative values of GCM predicted composite hydrological budget terms for the MSRB, and other major river basins in future model development efforts.

3.5. Multi-decadal trends

The multi-decadal trends simulated by each CMIP5 model of MSRB and annually averaged SM, P, E and R were analyzed to assess possible changes in hydrological quantities over the MSRB assuming continuation of current-day climate forcing. The time period analyzed was from the middle of the 20th century into the 21st century, utilizing the CMIP5 'historical' and 'rcp85' runs ('rcp45' for HadCM3). For this analysis, we annually average the monthly model output so that variance in the regressed series is solely due to year-to-year variability, rather than also including a component due to the peaks and troughs of the annual water cycle.

The values and twice the standard errors of the regression line slopes for the model simulations of SM, P, E and R are presented in Table 4. For six of the nine models (HadGEM-ES, CCSM4, HadCM3, GISS-E2-R, INM-CM4, and IPSL-CM5A-LR), the magnitude of projected SM anomaly trends is low, on the order of ~10 mm per century. The MRI-CGCM3 and GFDL-CM3 have larger trend magnitudes, of around 100 mm per century. The MIROC5 SM trends are near-zero up to around year 2010, then turn strongly negative after that at a rate of ~500 mm per century. This MIROC5 result is peculiar, both because the negative trend is so large and because it does not appear to be matched by the much more modest P, E and R trends produced after 2010. Overall, seven of the nine models show a negative SM anomaly trend (drying), while two (GFDL-CM3 and HadGEM2-ES) show a positive (moistening) trend.

In all models except HadGEM2-ES and INM-CM4, where missing values in the output do not allow determination, and MIROC5 before 2010, where trends are very small, the signs of trends in P, E and R are consistent, i.e. positive (negative) P trends accompany positive (negative) E and R trends. Changes in source and sink terms in the modeled hydrological balance therefore compensate one another in these models. For five of the models (CCSM4, GISS-E2-R, GFDL-CM3, HadCM3 and MRI-CGCM3) trends in P, E and R are positive, while for one model (IPSL-CM5A-LR) trends in P, E and R are negative. Of the five models with positive trends, only GFDL-CM3 shows a positive trend for SM anomaly alongside that for P, E, and R. The remaining four models with positive P, E and R trends (CCSM4, GISS-E2-R, HadCM3 and MRI-CGCM3) have negative SM anomaly trends.

Qian et al. (2007), using standard operational rain-gauge and surface meteorological observations to drive the CLM3 land-surface model, found positive trends in the MSRB for each SM, P, E and R over the time period 1948–2004. CMIP5 models with positive P, E and R trends are therefore consistent with the mid to late 20th century trends in these quantities found by Qian et al. (2007). Only the GFDL-CM3 model fully coincides with Qian et al. (2007) in having positive trends of SM, P, E and R. Long-term basin wide observations of SM, however, are not available to confirm the 20th century SM increase within the MSRB found in the Qian et al. (2007) CLM3 simulations. Also, as seen in Sections 3.1 and 3.4, the GFDL-CM3 model composite SM anomaly annual cycle compares less favorably with GRACE data compared to other models, and its predicted long-term average precipitation over the MSRB is about 20–30% higher than observed.

In Fig. 5 we depict for each CMIP5 model the differences in average SM, P, and P–E over three time periods, 1975–2000, 2001–2025, and 2026–2050, compared to the 1950–1974 average. On the vertical axes of the plots, the models are arranged from top to bottom according to increasing MAEs versus GRACE in the comparison of model composite annual TWS cycle presented in Table 2. By inspecting the plots, some correspondence of the simulated trends with model MAE scores can be inferred. The best performing models in terms of MAE tend to project drying SM but only slight

Table 4
Linear trends (mm per century) of SM, P, E and R computed from output of CMIP5 models. The time period utilized in regression calculation is given in parenthesis next to the model name. Twice the standard error of the slope value given in parenthesis next to slope value. Trends that are significant at the 95% confidence level are bolded.

Model (time period)	Regression line slopes (mm/century)			
	SM	P	E	R
CCSM4 (1936–2100)	**−15.9 (9.5)**	**3.5 (2.6)**	**2.8 (1.2)**	0.7 (0.7)
INM-CM4 (1936–2100)	**−27.1 (11.1)**	–[b]	0.3 (1.4)	−0.5 (1.2)
MIROC5[c] (2020–2100)	**−496 (62)**	−3.3 (8.6)	1.8 (2.3)	**−4.6 (3.1)**
HadGEM2-ES (1935–2100)	**23.6 (5.1)**	**3.1 (1.5)**	**1.7 (0.8)**	–[a]
HadCM3 (1936–2035)	**−9.3 (4.9)**	1.6 (1.7)	**1.4 (0.6)**	0.2 (1.2)
MRI-CGCM3 (1950–2100)	**−128 (10)**	**9.0 (2.1)**	**7.4 (0.6)**	1.7 (1.7)
GFDL-CM3 (1950–2050)	**94 (34)**	**15.8 (6.0)**	**13.1 (2.3)**	2.5 (2.6)
GISS-E2-R (1926–2050)	**−12.4 (8.8)**	**4.4 (2.1)**	**1.9 (0.7)**	**1.3 (0.9)**
IPSL-CM5A-LR (1936–2100)	**−24.4 (3.8)**	**−3.7 (2.5)**	−0.9 (0.9)	**−2.7 (1.6)**

[a] Runoff regression not performed for HadGEM2-ES, as 'mrro' output is not archived.
[b] Precipitation regression not performed for INM-CM4 due to periods of missing values in the output in the model 'historical' run.
[c] The MIROC5 SM slope before 2020 is slightly positive, then turns sharply negative afterwards.

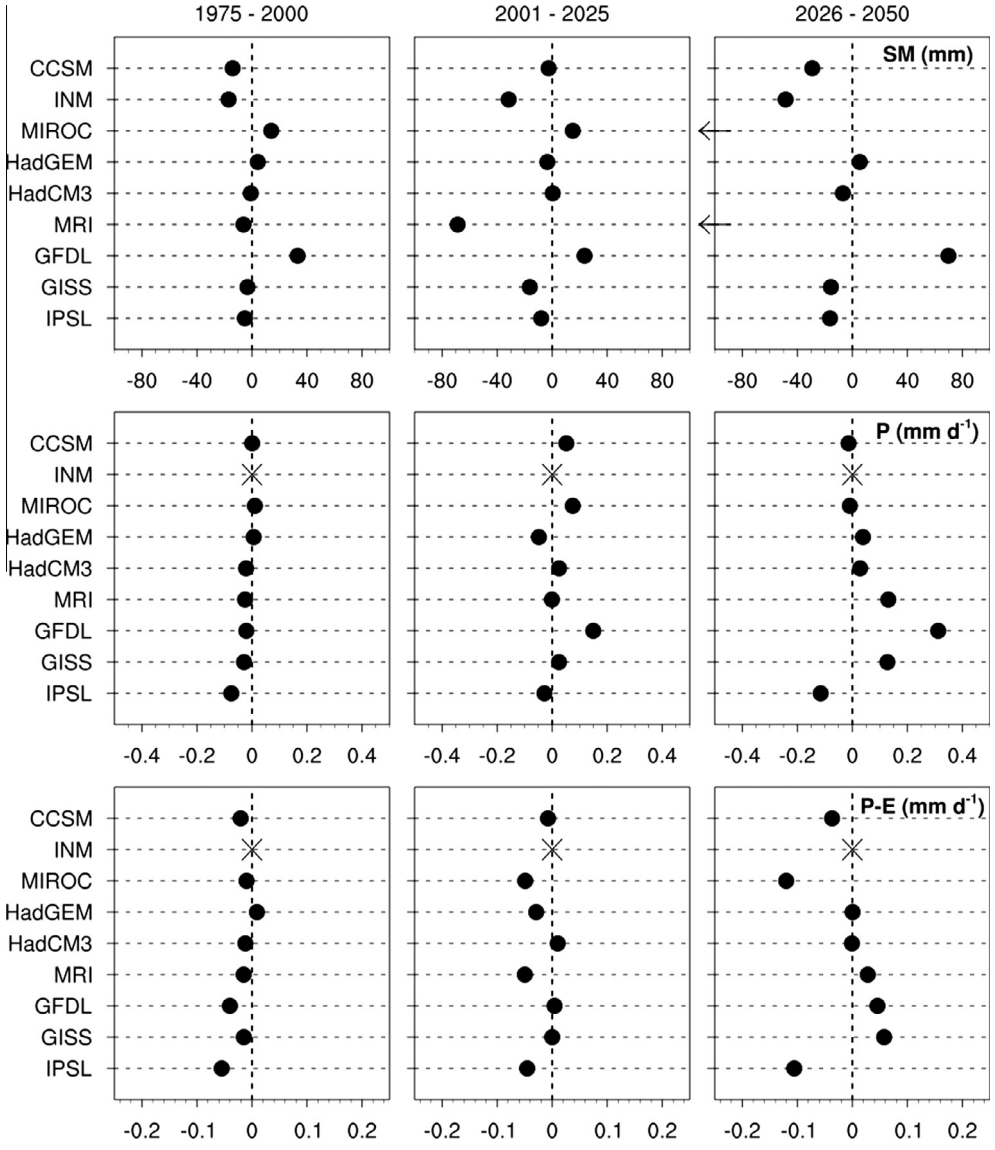

Fig. 5. Differences between 1975–2000, 2001–2025 and 2026–2050 period average values of SM (mm), P (mm day^{-1}) and P–E (mm day^{-1}) and their respective 1950–1974 averages. Models arranged from top to bottom on vertical axis according to increasing MAE of their composite annual TWS relative to GRACE TWS (Table 2). Left arrows on the SM plots for MIROC5 and MRI-CGCM3 indicate values less than −100 mm. 'X's on the P and P–E plots indicate that no calculation was made for INM-CM4 due to missing values in the archived 'pr' output file.

increasing P into the 21st century. The different magnitudes in the drying, especially noting the large drying projected by MIROC5 and MRI-CGCM3, as well as the more scattered response in P–E, however, indicates varying responses of the models in terms of details.

In summary, in spite of the better agreement of CMIP5 models with GRACE TWS anomaly data in representing the MSRB-average composite annual hydrological cycle, and the decreased variability among CMIP5 model predictions compared to CMIP3 predictions in simulating the cycle, model consensus is still broadly indeterminate on the direction of multi-decadal hydrological trends assuming continuation of current-day climate forcing. While there appears to be a drying consensus in SM trends among models, P, E and R trends do not show a consensus. Models with relatively low MAE values compared to GRACE (CCSM4, HadCM3, HadGEM2-ES and GISS-E2-R) simulate increasing P, E and R assuming continuation of current climate forcing. While this may indicate a consensus of sorts, more time, data and investigation are needed to confirm that decreasing SM coincides with increasing P, E and R. Clearly, prediction of hydrological quantities is complex, and depends on the time scale (annual vs. long-term decadal) of the dynamics considered.

4. Conclusion

We used measurements of terrestrial water storage (TWS) inferred from GRACE satellite observations to evaluate the TWS output and associated hydrological budget quantities composite-averaged over the Mississippi River Basin (MSRB) and over the ten-year period 2003–2012 from a subset of CMIP models. Modeled TWS was estimated as the sum of CMIP archived depth-integrated soil moisture and snow water. We then compared the model soil moisture spatial patterns across the MSRB against NARR reanalysis fields, and compared modeled composite precipitation, evapotranspiration and runoff to long-term observations and GRACE data. We finally investigated model simulated mid-20th to 21st century hydrological trends. The above analysis was based on spliced CMIP5 'historical' and 'rcp85' simulations, and CMIP3 A2 simulations.

Improvement was found in the simulations for most CMIP5 models compared to their earlier CMIP3 versions in predicting the composite annual cycle of TWS anomalies. The mean-absolute error (MAE) of monthly values in the multi-model mean ten-year composite annual cycle relative to the GRACE TWS composite improved from around 10 mm in the CMIP3 subset to around 3 mm in the CMIP5 subset. Intermodel variability of MAE values also reduced in CMIP5 compared to CMIP3 as a consequence of this improvement. These improvements appear associated with increased horizontal resolution and improved hydrological parameterizations in CMIP5 models compared to earlier CMIP3 versions. Quantitative uncertainty analysis, accounting for GRACE measurement error and uncertainties due to the use in some models of a single run, rather than an ensemble mean, to compare to GRACE indicates that intermodel differences of MAE values up to around 5 mm are likely insignificant. From analysis of one model (CCSM4), MAEs of GRACE versus modeled TWS estimated as the sum of soil moisture and snow water, and GRACE versus modeled TWS also including surface water and explicit groundwater in estimating modeled TWS, are different by amount 4 mm. The improvement in the multi-model mean MAE versus GRACE from around 10 mm in CMIP3 to around 3 mm in CMIP5 exceeds these uncertainty values, and is therefore likely significant.

Using GRACE observations, a value $\Delta TWS \approx 120$ mm is calculated for the peak to trough difference in the composite TWS annual cycle for the MSRB. This quantifies the average annual amount of water cycling into and out of the soil over the basin. Combining this with long-term river discharge observations for long-term average annual R, an observed value $\Delta TWS/R \approx 0.6$ is determined for the basin. Model analogues of ΔTWS were computed using the CMIP model soil moisture plus snow water composites. Model predicted values of ΔTWS and $\Delta TWS/R$ generally improved from CMIP3 to CMIP5 in comparison with the observed values based on GRACE ΔTWS (coinciding with the improved MAE scores), with the multi-model mean CMIP5 value closely matching the observed values. Intermodel variability of composite values of hydrological forcing terms, however, is large. In particular, CMIP5 models appear grouped according to some that developed increased ΔTWS as a result of increased composite annual average P and others that developed increased ΔTWS as a result of decreased R. The changes from CMIP3 to CMIP5 in P and R for many of the analyzed models were outside observational ranges. More research is therefore needed to better calibrate GCM predictions of the composite hydrological cycle of the MSRB, and likely other major river basins as well. Observed measures of soil water infiltration and retention from GRACE TWS anomalies, through ΔTWS, $\Delta TWS/R$ or other metrics, appear useful for such model development efforts.

In spite of the improved predictions by CMIP5 models of the average annual hydrological cycle over the MSRB, multi-decadal hydrological trends for the MSRB simulated by the models from the mid-20th to 21st century subject to current-day climate forcing remain variable. Although a majority of the models (five of the nine investigated) project decreased SM on the order of 10 mm per century in response to the forcing, two others predict decreases on the order of 100 mm per century, and two others predict increased SM. Responses of P, E and R also vary appreciably among the model simulations.

Future work can shed light on these findings. Finer spatial analysis using GRACE data can be performed to assess GCM performance for sub-regions within the MSRB. This is possible using current GRACE Release 5 as well as future GRACE-FO data, which are suitable for analysis down to horizontal scales of around 50,000 km^2. Finer spatial analysis is envisioned using data from future launches beyond GRACE-FO (Famiglietti and Rodell, 2013). A longer time record of GRACE data, as more of its retrievals become available, will enable evaluation of the TWS storage term in the hydrological balance on longer time scales, which will help validate GCM produced long-term hydrological trends in TWS.

Acknowledgements

The authors thank Mr. Gary Strand at the NCAR Earth System Laboratory for making available and assisting us in understanding CCSM4 model explicit groundwater and surface water grids from http://www.earthsystemgrid.org. We also thank Ms. Kelly McDonnell at San Jose State University Department of Meteorology and Climate Science for her technical assistance in drafting this manuscript.

References

Anderson, O.B., Seneviratne, S.I., Hinderer, J., Viterbo, P., 2005. GRACE-derived terrestrial water storage depletion associated with the 2003 European heat wave. Geophys. Res. Lett. 32, L18405. http://dx.doi.org/10.1029/2005GL023574.

Cai, X., Yang, Z., David, C.H., Niu, G., Rodell, M., 2014. Hydrological evaluation of the NOAH-LP model for the Mississippi River Basin. J. Geophys. Res. 119, 23–38.

CLIVAR, 2011. Data and Bias Correction for Decadal Climate Predictions. CLIVAR Publication Series 150, International CLIVAR Project Office, 4 pp.

Collow, T.W., Robock, A., Basara, J.B., Ilston, B.G., 2012. Evaluation of SMOS retrievals of soil moisture over the central United States with currently available in situ observations. J. Geophys. Res. 117, D09113. http://dx.doi.org/10.1029/2011JD017095.

Cox et al., 1999. The impact of new land surface physics on the GCM simulation of climate and climate sensitivity. Clim. Dyn. 15, 183–203.

CSR (Center for Space Research), 2014. GRACE Mission Data Flow. <http://www.csr.utexas.edu/grace/asdp.html>.
Dufresne et al., 2013. Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5. Clim. Dyn. 40, 2123–2165.
Famiglietti, Rodell, 2013. Water in the balance. Science 340, 1300–1301.
Famiglietti, J.S. et al., 2011. Satellites measure recent rates of groundwater depletion in California's Central Valley. Geophys. Res. Lett. 38, L03403.
IPCC (Intergovernmental Panel on Climate Change), 2013. Climate Change 2013: The Physical Science Basis. <http://www.ipcc.ch/report/ar5/wg1/#.UorSO7l3vIU>.
JPL (Jet Propulsion Laboratory), 2013. GRACE Monthly Mass Grids – Land. <http://grace.jpl.nasa.gov/data/gracemonthlymassgridsland/>.
Kim, H., Yeh, P.J.F., Oki, T., Kanae, S., 2009. Role of rivers in the seasonal variations of terrestrial water storage over global basins. Geophys. Res. Lett. 36, L17402.
Landerer, F.W., Swenson, S.C., 2012. Accuracy of scaled GRACE terrestrial water storage estimates. Water Resour. Res. 48, W04531–W04541. http://dx.doi.org/10.1029/2011WR011453.
Lawrence, D.M. et al., 2011. Parameterization improvements and functional and structural advances in version 4 of the Community Land Model. J. Adv. Model. Earth Syst. 3, 1–28.
Li, H., Robock, A., Wild, M., 2007. Evaluation of Intergovernmental Panel on Climate Change Fourth Assessment soil moisture simulations for the second half of the twentieth century. J. Geophys. Res. 112, D06106. http://dx.doi.org/10.1029/2006JD007455.
Lu, E., Takle, E.S., Minoj, J., 2010. The relationships between hydrological and climatical changes in the Upper Mississippi River Basin: a SWAT and multi-GCM study. J. Hydrometeorol. 11, 437–451.
Martin et al., 2011. The HadGEM2 family of met office unified model climate configurations. Geosci. Mod. Dev. 4, 723–757.
Meehl, G.A., Covey, C., Delworth, T., Latif, M., Mcavaney, B., Mitchell, J.F.B., Stouffer, R.J., Taylor, K.E., 2007. The WCRP CMIP3 multimodel dataset – a new era in climate change research. Bull. Am. Meteor. Soc. 88, 1383–1394.
Mesinger, F. et al., 2006. North American regional reanalysis. Bull. Am. Meteor. Soc. 87, 343–360.
Miguez-Macho, G., Li, H., Fan, Y., 2008. Simulated water table and soil moisture climatology over North America. Bull. Am. Meteorol. Soc. 89, 663–672.
Milly, P.C.D., 2005. Trends in the Water Budget in the Mississippi River Basin, U.S.G.S Fact Sheet. <http://pubs.usgs.gov/fs/2005/3020/>.
Milly, P.C., Malyshev, S., Shevliakova, E., Dunne, K.A., Findell, K.L., Gleeson, T., Liang, Z., Phillipps, P., Stouffer, R.J., Swenson, S.C., 2014. An enhanced model of land water and energy for global hydrologic and earth-system studies. J. Hydrometeorol. 15, 1739–1761.
Milly, P.C.D., Dunne, K.A., Vecchia, A.V., 2005. Global pattern in trends of streamflow and water availability in a changing climate. Nature 438, 347–350.
Music, B., Caya, D., 2007. Evaluation of the hydrological cycle over the Mississippi River Basin as simulated by the Canadian Regional Climate Model (CRCM). J. Hydrometeor. 8, 969–988.
Niu, G., Yang, Z., Dickinson, R.E., Gulden, L.E., Su, H., 2007. Development of a simple groundwater model for use in climate models and evaluation with the Gravity Recovery and Climate Experiment data. J. Geophys. Res. 112, D07103.
Oleson, K.W. et al., 2008. Improvements to the Community Land Model and their impact on the hydrological cycle. J. Geophys. Res. 113, G01021.
Oleson, K.W. et al., 2010. Technical Description of Version 4.0 of the Community Land Model (CLM), NCAR/TN-478+STR.
PCMDI (Program for Climate Model Diagnosis and Intercomparison), 2010. About the WCRP CMIP3 Multi-Model Dataset Archive at PCMDI. <http://www-pcmdi.llnl.gov/ipcc/about_ipcc.php>.
Pokhrel, Y., Hanasaki, N., Koirala, S., Cho, J., Yeh, P.J.F., Kim, H., Kanae, S., Oki, T., 2012. Incorporating anthropogenic water regulation modules into a land surface model. J. Hydrometeorol. 13, 255–269.
Pokhrel, Y., Fan, Y., Miguez-Macho, G., Yeh, P.J.F., Han, S., 2013. The role of groundwater in the Amazon water cycle: 3. Influence on terrestrial water storage computations and comparison with GRACE. J. Geophys. Res. 118, 3233–3244.
Prigent, C., Aires, F., Rossow, W.B., Robock, A., 2005. Sensitivity of satellite microwave and infrared observations to soil moisture at a global scale: relationship of satellite observations to in situ soil moisture measurements. J. Geophys. Res. 110, D07110. http://dx.doi.org/10.1029/2004JD005087.
Qian, T., Dai, A., Trenberth, K.E., 2007. Hydroclimatic trends in the Mississippi River Basin from 1948 to 2004. J. Clim. 20, 4599–4614. http://dx.doi.org/10.1175/JCLI4262.1.
Ramillien, G., Famiglietti, J.S., Wahr, J., 2008. Detection of continental hydrology and glaciology signals from GRACE: a review. Surv. Geophys. 29, 361–374. http://dx.doi.org/10.1007/s10712-008-9048-9.
Rodell, M., Famiglietti, J.S., 2002. The potential for satellite-based monitoring of groundwater storage changes using GRACE: the High Plains aquifer, Central US. J. Hydrol. 263, 245–256.
Rodell, M., Chen, J., Kato, H., Famiglietti, J.S., Nigro, J., Wilson, C.R., 2007. Estimating groundwater storage changes in the Mississippi River basin (USA) using GRACE. Hydrogeol. J. 15 (1), 159–166. http://dx.doi.org/10.1007/s10040-006-0103-7.
Schmidt et al., 2006. Present-day atmospheric simulations using GISS ModelE: comparison to in situ, satellite, and reanalysis data. J. Clim. 19, 153–192.
Schumann, G., Lunt, D.J., Valdes, P.J., de Jeu, R.A.M., Scipal, K., Bates, P.D., 2009. Assessment of soil moisture fields from imperfect climate models with uncertain satellite observations. Hydrol. Earth Syst. Sci. 13, 1545–1553.
Seneviratne, S.I., Corti, T., Davin, E.L., Hirschi, M., Jaeger, E.B., Lehner, I., Orlowsky, B., Teuling, A.J., 2010. Investigating soil moisture–climate interactions in a changing climate: a review. Earth-Sci. Rev. 99, 125–161.
Sperna-Weiland, F.C., Van Beek, L.P.H., Weerts, A.H., Bierkens, M.F.P., 2012. Extracting information from an ensemble of GCMs to reliably assess future global runoff change. J. Hydrol. 412–413, 66–75.
Swenson, S.C., Wahr, J., 2006. Post-processing removal of correlated errors in GRACE data. Geophys. Res. Lett. 33, L08402. http://dx.doi.org/10.1029/2005GL025285.
Taylor, K.E., Stouffer, R.J., Meehl, G.A., 2012. An overview of CMIP5 and the experimental design. Bull. Am. Meteor. Soc. 93, 485–498.
Volodin, E.M., Dianskii, N.A., Gusev, A.V., 2010. Simulating present day climate with the INMCM4.0 coupled model of the atmospheric and oceanic general circulations. Iz. Atmos. Ocean. Phys. 46, 414–431.
Wang, G., 2005. Agricultural drought in a future climate: results from 15 global climate models participating in the IPCC 4th assessment. Clim. Dyn. 25, 739–753.
Watanabe, M. et al., 2010. Improved climate simulation by MIROC5: mean states, variability, and climate sensitivity. J. Clim. 23, 6312–6335.
Yukimoto, S. et al., 2011. Meteorological Research Institute-Earth System Model Version 1 (MRI-ESM1): Model Description, TECHNICAL REPORTS OF THE METEOROLOGICAL RESEARCH INSTITUTE No.64 (Meteorological Research Institute, Japan).
Zaitchak, B.F., Rodell, M., Reichle, R.H., 2008. Assimilation of GRACE terrestrial water storage data into a land surface model: results for the Mississippi River Basin. J. Hydrometeorol. 9, 535–548.
Zeng, N., Yoon, J.-H., Mariotti, A., Swenson, S., 2008. Variability of basin-scale terrestrial water storage from a PER water budget method: the Amazon and the Mississippi. J. Clim. 21, 248–265.