Evaluating the hydrological consistency of satellite based water cycle components

Oliver López[1], Rasmus Houborg[1], Matthew F McCabe[1]

[1]Water Desalination and Reuse Center, Division of Biological and Environmental Sciences and Engineering, King Abdullah University of Science and Technology, Thuwal, 23955-6900, Saudi Arabia

Correspondence to: Oliver López (oliver.lopez@kaust.edu.sa)

Abstract. Advances in multi-satellite based observations of the earth system have provided the capacity to retrieve information across a wide-range of land surface hydrological components and provided an opportunity to characterize terrestrial processes from a completely new perspective. Given the spatial advantage that space-based observations offer, several regional-to-global scale products have been developed, offering insights into the multi-scale behaviour and variability of hydrological states and fluxes. However, one of the key challenges in the use of satellite-based products is characterizing the degree to which they provide realistic and representative estimates of the underlying retrieval: that is, how accurate are the hydrological components derived from satellite observations? The challenge is intrinsically linked to issues of scale, since the availability of high-quality in-situ data is limited, and even where it does exist, is generally not commensurate to the resolution of the satellite observation. Basin-scale studies have shown considerable variability in achieving water budget closure with any degree of accuracy using satellite estimates of the water cycle. In order to assess the suitability of this type of approach for evaluating hydrological observations, it makes sense to first test it over environments with restricted hydrological inputs, before applying it to more hydrological complex basins. Here we explore the concept of hydrological consistency, i.e. the physical considerations that the water budget impose on the hydrologic fluxes and states to be temporally and spatially linked, to evaluate the reproduction of a set of large-scale evaporation (E) products by using a combination of satellite rainfall (P) and Gravity Recovery and Climate Experiment (GRACE) observations of storage change, focusing on arid and semi-arid environments, where the hydrological flows can be more realistically described. Our results indicate no persistent hydrological consistency in these environments, suggesting the need for continued efforts in improving satellite observations, particularly for the retrieval of evaporation, and the need to more directly account for anthropogenic influences such as agricultural irrigation into our large scale water cycle studies.

1 Introduction

Progress in satellite-based observations of the Earth system has enabled the characterization of land surface hydrological components and an improved representation of terrestrial processes (Famiglietti et al., 2015). Dedicated space missions such as the Gravity Recovery and Climate Experiment (GRACE) (Tapley et al., 2004b), the Global Precipitation Measurement

Mission (GPM) (Hou et al., 2014) and a suite of microwave-based soil moisture platforms (Liu et al., 2012), represent important efforts that have contributed to these advances. Considering the spatial advantage that space-based observations have over ground-based measurements, there has been a proliferation of regional-to-global scale data products, providing knowledge on the multi-scale behaviour and patterns of hydrological states and fluxes. However, one of the key challenges

5 of space-based remote sensing is to characterize the degree to which these products represent realistic estimates of the underlying variables they attempt to retrieve. Inherent to this challenge is the issue of scale, a consequence of both a lack of abundant high-quality in-situ data and the fact that there is an inevitable scale mismatch between these measurements (McCabe et al. 2006).

10 A crucial task that is required to address these questions is to evaluate the hydrological consistency amongst these different hydrological products. Hydrological consistency refers to the spatial and temporal match that must exist between individual observations of hydrological states and fluxes based on physical considerations. For example, cloud detection can be used to validate precipitation events (Milewski et al., 2009); soil moisture changes should closely match precipitation anomalies; changes in atmospheric related variables such as humidity and sensible heat flux should correspond to the soil moisture state

15 (McCabe et al., 2008); and, in a larger scale, the spatial distribution and timing of water storage anomalies should be closely related to precipitation anomalies. In principle, in regions where runoff is low,independent estimates of water storage, precipitation and evaporation should provide a physically-based closure of the water budget. However, even excluding the uncertainties inherent in the modeling and retrieval of these variables from satellite data, the complexities of land-surface dynamics, conditions and residence times also make it challenging to apply in practice. A number of studies have evaluated

20 the water budget closure of large basins across different regions of the world using either satellite observations alone (Sheffield et al., 2009) or via a combination of satellite observations and data assimilation (Pan and Wood, 2006; Pan et al., 2008; Sahoo et al., 2011; Pan et al, 2012). While the objective of Sheffield et al. (2009) was to evaluate the water budget closure (by comparing the residual of the water budget with measured streamflow values), the remaining studies mostly aim to provide merged/constrained estimates of the water cycle components. When considering these earlier contributions, there

25 remains a need to determine whether the goal of hydrological consistency can be realistically achieved using currently available satellite products. In examining this question, it makes sense to focus on regions with relatively simple water budgets, e.g. without snow, dense vegetation or river components, as they represent natural laboratories within which the evaluation of such large-scale products and the concept of hydrological consistency could be reasonably undertaken.

30 In this study, we focus our analysis on four large river basins, based on their resemblance to an ideal water budget test-case, to examine the hydrological consistency of satellite observations of water storage, precipitation and evaporation. Total water storage, i.e. the summation of groundwater, soil moisture, snow, surface water, ice and biomass, were derived from anomalies in the gravity field from GRACE satellites (Tapley et al., 2004b). GRACE water storage estimates have been used in a myriad of studies exploring the indirect groundwater responses across many different spatial and temporal scales

(Swenson et al., 2008; Rodell et al., 2009; Famiglietti et al., 2011; Sun, 2013; Voss et al., 2013). The accuracy of GRACE terrestrial water storage anomalies (TWSA) is related to the number of degrees to which the gravity field is solved for in spherical harmonics (Swenson and Wahr, 2002) and an approximate global averaged accuracy of 20 mm/month has been previously proposed (Wahr and Velicogna, 2006).

Terrestrial evaporation (E), comprising the sources of soil and canopy evaporation together with plant transpiration, serves as a key component in our analysis. Unlike microwave or radiative emissions from the surface or atmosphere that can be used to inform upon soil moisture, surface temperature or rainfall, evaporative fluxes provide no directly observable trace that can be detected from satellites and are instead estimated through interpretive or empirical models (Jimenez et al., 2011; Ershadi

10 et al., 2014). Recently, several of these models have been used to estimate global scale evaporation by combining satellite observations of surface variables with meteorological and other ancillary data (McCabe et al., 2016; Miralles et al., 2016). While some of these models have relatively simple parameterizations (Fisher et al. 2008), they still require considerable amounts of input data that may not be available everywhere at a global scale (Jimenez et al. 2011). Available ground observations are routinely used to calibrate and evaluate these models. However, the large-scale implementation of such

15 approaches is inevitably constrained by the lack of distributed and representative in-situ networks with which to comprehensively assess simulations. Recent evaluation efforts have shown that no single evaporation product consistently outperforms any other, whether applied at local or global scales (Ershadi et al., 2014; Miralles et al., 2016). Recognizing this uncertainty, we employ three different evaporation products to account for differences in evaporation output deriving from the different models and/or input data.

Overall, the purpose of this paper is to evaluate the hydrological consistency of independent satellite observations over basins where it is assumed that the water cycle system is relatively simple. Through this analysis we expect to determine whether the concept of hydrological consistency can be employed in regions with more complicated water cycle systems to aid in the validation of evaporation and other hydrological data sets. A secondary objective is to determine the impact that

25 the choice of different evaporation products has on the analysis: is hydrological consistency achieved with a particular product, or is the disagreement between evaporation products significant enough to impact the study? Furthermore, if the hydrological consistency approach is not achievable, what does this say about the retrieval accuracy of these independent observations of the hydrological cycle? Section 2 describes the data sources, including a brief description of the global evaporation products used. Section 3 presents in detail the methodology used to evaluate the hydrological consistency based

30 on a spherical harmonic analysis. The results in Section 4 show the spatial and temporal behavior of the correlations between water storage anomalies and P-E, while the implications of these are discussed in greater detail in Section 5. Finally, concluding remarks are provided in Section 6.

2 Data sources and study regions

A range of globally distributed large scale data sets derived primarily from satellite observations were used in this analysis. The study period, encompassing the years between 2003-2011, was based upon the availability of GRACE data and several evaporation products. In the following paragraphs we briefly describe the sources and nature of the data used in this
5   contribution.

2.1 GRACE water storage anomalies

Water storage anomalies (2003-2011) were derived from GRACE (release 05) monthly spherical harmonic coefficients representing the gravity field, processed at the University of Texas Center for Space Research (UTCSR). The degree 2
10   coefficients were replaced by satellite laser ranging values calculated by Cheng et al. (2011) and Cheng et al. (2013). The coefficients were filtered using a de-striping technique to remove correlated errors in the coefficients that would otherwise obscure the signal (Swenson and Wahr, 2006; Duan et al., 2009). An isotropic (Gaussian) filter was then used to remove random errors (Swenson and Wahr, 2002). While the effect that the de-striping filter has on the true geophysical signal is not known a priori, an indirect measure can be obtained by applying the filter to a synthetic water storage variation from a land
15   surface model (LSM). This method is used to obtain scaling factors for GRACE data in order to restore the signal (Landerer and Swenson, 2012). Long et al. (2015) evaluated the impact of different land surface models on the scaling factor and showed that the impact was greatest in arid regions. To avoid this potential factor of uncertainty in our study, which is focused on arid regions, we instead transformed the other water cycle components into spherical harmonics. The effect of the filters is therefore incorporated directly into the study by applying the filters to the various hydrological components in
20   spherical harmonics (see Section 3.1).

2.2 Evaporation products

Several satellite-based evaporation products have been developed over the last decade, based on a range of modeling schemes (Mu et al., 2011; Leuning et al. 2008; Miralles et al., 2011). Given the importance of evaporation within studies of the global energy and water cycle, considerable effort has been directed towards accurately reproducing its spatial and
25   temporal variability, with comprehensive reviews of various approaches to do this provided by Kalma et al. (2008) and Wang and Dickinson (2012). Here we employ a range of global evaporation datasets, which are described briefly in the following paragraphs and summarized in Table 1. To ensure consistency with the GRACE data, the evaporation products were aggregated from daily (or 8-daily in the case of MOD16) to monthly estimates, centered on the dates specified in the GRACE monthly gravity field solutions. In the aggregation from daily to monthly data, pixels that presented missing
30   data for more than 20% in a given month were not included in the calculation.

2.2.1 MOD16

Cleugh et al. (2007) developed an algorithm for large-scale evaporation monitoring based on the Penman-Monteith (PM) equation, using meteorological forcing data and a surface resistance linearly modeled through remotely sensed Leaf Area Index (LAI) and Normalized Difference Vegetation Index (NDVI), as measured by the MODerate resolution Imaging
5 Spectroradiometer (MODIS). Improvements to this approach (Mu et al., 2007; Mu et al., 2011) led to the development of the MODIS Global Evapotranspiration product (MOD16). In MOD16, the linearization of the surface resistance is specified for each biome separately via a look-up table, with the evaporation calculated for daytime and nighttime conditions. Other improvements include: soil heat flux calculation, distinction of dry and wet canopy as well as moist and wet soil, and improvements to the aerodynamic resistance. The product includes transpiration and evaporation from soil and wet canopy,
10 as well as total evaporation calculated as the sum of these three components. Potential evaporation (calculated using a Priestley-Taylor based formulation) is also included to monitor environmental water stresses and droughts (Mu et al., 2011).

2.2.2 CSIRO-PML

In parallel to the PM-Mu model, Leuning et al. (2008) introduced improvements to the Cleugh et al. (2007) algorithm, resulting in the Penman-Monteith-Leuning (PML) model. An important new feature of the PML approach was a biophysical
15 algorithm for the calculation of canopy conductance. Zhang et al. (2010) later introduced an approach to model the spatial variability of two parameters in the PML model (maximum stomatal conductance and the ratio of actual to potential evaporation at the soil surface) that were previously held constant, leading to the development of a global daily evaporation product at 0.25-degree resolution. In this manuscript, the global product is referred to as the Commonwealth Scientific and Industrial Research Organisation (CSIRO)-PML product.

2.2.3 GLEAM (v2A)

The Global Land Evaporation: the Amsterdam Methodology (GLEAM) (Miralles et al., 2011) is a satellite data based model developed to estimate evaporation at a global scale. In this approach, rainfall interception loss is evaluated using an analytical model (Gash, 1979). Next, satellite observations of surface soil moisture are assimilated using a Kalman filter assimilation approach to estimate the moisture profile over several soil layers. The evaporation from bare soil, short
25 vegetation and tall canopy vegetation is calculated using the Priestley-Taylor equation and a stress factor based on environmental conditions.

2.3 Precipitation data

Global daily precipitation (P) estimates derived from multi-satellite observations for the period 2003-2011 were obtained from the Global Precipitation Climatology Project (GPCP) (Huffman et al., 2001), the official World Climate Research
30 Program (WCRP) Global Energy and Water Cycle Exchanges (GEWEX) product. The GPCP product is a merged

precipitation analysis combining information from microwave, infrared, and sounder data observed by a constellation of international precipitation-related satellites (Huffman 1997; Huffman 2001; Adler 2003). The estimates from microwave and infrared data are based on the Threshold-Matched Precipitation Index (TMPI). The combined satellite-based product is corrected by rain gauge analysis where data is available. Over many areas of the world, the GPCP product represents one of

5   the best available sources of precipitation data (Crow, 2007; Miralles et al., 2011). In this study, we used the daily product (GPCP 1DD) and converted daily values to monthly estimates, centered on the dates provided in GRACE monthly gravity field solutions. Pixels were assigned as missing data when more than 20% of the month was missing (on a per-pixel basis).

2.4 Runoff data

Although not considered in the evaluation of hydrological consistency, runoff data is compared to precipitation and

10   evaporation data in order to evaluate whether the assumption of a simple water budget is valid in the study regions. Surface runoff, sub-surface runoff and snowmelt were derived from the NOAH land surface model included in the Global Land Data Assimilation System (GLDAS) (Rodell et al., 2004). GLDAS uses global satellite and ground-based observational products to obtain optimal estimates of land surface states and fluxes from land surface models using data assimilation techniques. The version of the product used in this study (GLDAS-2.0) is forced with meteorological data from the Princeton University

15   forcing data set (Sheffield et al., 2006) and is available at 1-degree resolution from 1948 to 2010.

2.5 Study regions

The study regions were chosen mainly on climate classification, with river basins that covered regions with a predominantly arid climate preferentially selected. We employed a Köppen classification map generated using data sets from the Climatic Research Unit and the Global Precipitation Climatology Centre up to 2006 (Kottek et al., 2006). The climate criteria was to

20   select basins with more than 50% areal extent containing any of the arid Köppen climates (BWk, BWh, BSk or BSh). The basins were selected from a set of 405 globally distributed river basins provided by the Global Runoff Data Centre (GRDC), derived from flow direction data of the HYDRO1k Elevation Derivative Database developed at the U.S. Geological Survey (USGS). Secondary considerations for selecting basins from the GRDC data set were based on size, geographical location and amplitude and trends in the water storage variations. We selected four basins as focus regions of study: the Colorado

25   River basin in North America, Niger basin in Africa, Aral basin in Asia and the Lake Eyre basin in Australia (Figure 1).

Figure 2 shows the spatially averaged hydrological fluxes over the study basins, including the sum of surface, subsurface runoff and snowmelt runoff (Q) derived from the GLDAS NOAH version 2 monthly product (Rodell et al., 2004). Q is included in these figures to compare the extent to which a major assumption of the study, i.e. a simple water budget, is met

30   across each of the study regions. Although pre-dominantly arid with a combination of hot arid desert and cold arid steppe climate classifications, both the Colorado River basin (CRB) and the Aral Sea basin (ASB) contain a snow component. Snowmelt in these two regions plays an important role in the water cycle, particularly in the delivery and redistribution of

water to other areas of the basin. Therefore, hydrological consistency might not be satisfied completely in these regions using our simple water budget assumption for some periods. Similarly, the Niger River basin (NRB) also has a runoff component as important as evaporation, but we assume that it will not affect the spatial distribution of water storage anomalies. Finally, in the Lake Eyre basin (LEB) we expect that the limited and sporadic runoff component will not have a

5   significant effect on the hydrological consistency analysis.

3 Methodology

In order to provide a meaningful spatial evaluation of the hydrological consistency between the data sets (i.e. at sub-basin scale), the analysis was carried out in spherical harmonics. Furthermore, the effects of the de-striping filter are accounted for directly instead of relying on a land surface model, the choice of which can severely impact the results of our analysis in arid

10   regions (Long et al., 2015). In this section, we present a detailed account of how the transformation was carried out, as well as how the actual evaluation of hydrological consistency is performed in spherical harmonics.

3.1 Spherical harmonic analysis of evaporation and precipitation data sets

As noted earlier, in order to provide a comparison at a sub-basin scale and to incorporate the effects of the de-striping filter for GRACE data into the study, the various components of the water cycle examined here were transformed into spherical

15   harmonics. Any continuous function $f(\theta,\phi)$ defined on the surface of a sphere, in this case, the global evaporation (E) and precipitation (P) data sets, can be approximated by a finite set of spherical harmonic coefficients C_{lm} and S_{lm} (Wang et al., 2006; Swenson and Wahr, 2002), following Eq. (1):

$$f(\theta,\phi) \approx \sum_{l=0}^{l_{max}} \sum_{m=0}^{l} \widetilde{P_{lm}}(\cos(\theta))(C_{lm}\cos(m\phi) + S_{lm}\sin(m\phi)), \tag{1}$$

where $\tilde{P}_{lm}(\cos\theta)$ is the normalized associated Legendre functions, θ corresponds to the colatitude (the complementary angle

20   to the latitude), ϕ to longitude and l_{max} is the truncation degree. The approximation is closer to the true value of f for large values of l_{max}, with an approximate relation between l_{max} and spatial scale λ given by $\lambda \approx 20{,}000$ km^2/ l_{max}. GRACE coefficients are available up to degree 60 every month, i.e. with a horizontal length scale of $20{,}000/60 \approx 330$ km.

The spherical harmonic analysis refers to the process of solving equation 1 for a set of coefficients C_{lm} and S_{lm}. Several

25   computational packages are available to perform this type of analysis. In this study, we used a FORTRAN program developed by Wang et al. (2006), which is suited for regularly gridded regional and/or global non-smooth data sets. The program can also perform spherical harmonic synthesis, which is the inverse transformation (i.e. from coefficients to spatial data). Figure 3 presents an example of the transformation based on the gridded CSIRO-PML data for April 2003.

30   Because all data sets are evaluated up to the same degree l_{max}, any differences due to the mismatch in the resolution of the products are eliminated after spherical harmonic analysis and synthesis. After this process, we generated three P−E anomaly

data sets, i.e. one for each E data set. Next, we applied the de-striping and Gaussian filters to account for the effect that these have in GRACE TWSA data.

3.2 Regional spherical harmonic analysis

The computed spherical harmonic coefficients so far are global (e.g. Figure 3). In order to evaluate the hydrological

5   consistency of the study regions (Figure 1), the data needed to be masked for the particular study regions. In Swenson and Wahr (2002), an exact averaging kernel is defined as a function with a value of 1 inside the boundaries of a region and 0 outside. To isolate the GRACE signal, an approximated averaging kernel was computed in spherical harmonics and convolved with the Gaussian filter in order to obtain a spatially averaged value of the TWSA (at a basin-scale). In this study, we instead compute the spherical harmonic analysis of the product of the global data sets (e.g. TWSA or P-ET) with the

10   averaging kernel following Eq. (2):

$$f^b(\theta, \phi) = f^g(\theta, \phi)\vartheta(\theta, \phi), \tag{2}$$

where $f^b(\theta,\phi)$ is the isolated regional data, $f^g(\theta,\phi)$ is the global data set and $\vartheta(\theta,\phi)$ is the averaging function. The relation in spherical harmonics is given by Eqs. (3-4):

$$\Delta S^b = \sum_{j_1,m_1} \sum_{j_2,m_2} \Delta S^g_{j_1,m_1} B_{j_2,m_2} Q^{jm}_{j_1m_1j_2m_2}, \tag{3}$$

15   $$Q^{jm}_{j_1m_1j_2m_2} = \sqrt{\frac{(2j_1+1)(2j_2+1)}{4\pi(2j+1)}} C^{j0}_{j_10j_20} C^{jm}_{j_1m_1j_2m_2}, \tag{4}$$

where $C^{jm}_{j1m1j2m2}$ are the Clebsch-Gordan coefficients (Martinec, 1989). We used the program developed by Martinec (1989) to mask the three global P-ET data sets (as well as GRACE data) over the four study regions (Figure 1).

3.3 Evaluating spatial agreement in the spherical harmonics of two data sets

The spatial agreement between two data sets can be evaluated using spherical harmonic coefficients by computing the degree

20   correlation measure (Arkani-Hamed, 1998; Tapley et al., 2004a) following Eq. (5):

$$r_l = \frac{1}{\sigma_l^{(A)}\sigma_l^{(B)}} \sum_{m=0}^{l} (C^{(A)}_{lm}C^{(B)}_{lm} + S^{(A)}_{lm}S^{(B)}_{lm}), \tag{5}$$

where σ_l^2 is the degree variance given in Eq. (6):

$$\sigma_l^2 = \sum_{m=0}^{l} (C^2_{lm} + S^2_{lm}), \tag{6}$$

The degree correlation measure is computed for every degree (l), and therefore we can in principle evaluate the hydrological

25   consistency at different length scales (i.e. sub-basin variability). As noted earlier, GRACE data is limited in resolution by l_{max} =60, or to approximately 330 km. In practice however, the de-striping filter removes all coefficients larger than 40 and therefore we are limited to length scales of about 500 km and larger. The smallest region in this study is the Colorado River Basin, covering an area of about 640,000 km^2. Based on this area, we can set a limit for the approximate largest spatial scale relevant to our study as 800 km, corresponding approximately to degree 25.

4 Results

4.1 Assessing the consistency of evaporation products

As noted earlier, an objective of this work is to determine whether the choice of different evaporation products affects the hydrological consistency analysis i.e. is the disagreement between evaporation products significant enough to impact the outcomes of the study? A cursory examination of the evaporation data sets indicates that there are evident differences across the different products in each of the studied basins (see Figure 2). In general, MOD16 simulates lower flux estimates when compared against both CSIRO-PML and GLEAM, a feature that has been noted in a number of recent global intercomparison studies (McCabe et al., 2016; Michel et al., 2016; Miralles et al. 2016). There are also clear differences in terms of the variability in the temporal response of the models, although CSIRO-PML and GLEAM show a greater level of agreement in terms of amplitude and timing, if not in absolute values. For example, during the wet period of 2004-2005 in the Colorado river basin (CRB), the response to precipitation reflected in MOD16 was far more rapid than either CSIRO-PML or GLEAM displayed. Of some concern is that CSIRO-PML is larger than precipitation during much of the study period in both the Colorado and the Lake Eyre basins, immediately constraining any type of hydrological consistency analysis. In the Niger river basin (NRB), there is more consistent agreement between the evaporation products, indicating greater confidence in the retrievals of evaporation in this region. For the Aral Sea basin (ASB), the discrepancies in E are similar to those obtained for the CRB region, with an obvious phase shift in CSIRO-PML and GLEAM observed relative to MOD16. This may reflect complexities in evaporation modeling due to the intermixed climate zones in the region caused by differences in land surface parameters. In the Lake Eyre basin (LEB), there are differences in amplitude but not in the temporal behavior of E.

Overall, from even just a qualitative perspective, there are clear challenges in developing a hydrological closure approach over these comparatively simple basins. While it is not the intent of this current work to explore the error characterization of these different models, the techniques being used to evaluate product consistency do provide some insight into retrieval quality: at least relative to the other hydrological products being compared against. These ideas are explored more quantitatively in the following sections.

4.2 Basin scale assessment

A key objective of this analysis is to assess the hydrological consistency between discrete components of the water cycle (see Figure 2). To do this, we examined the spatial and temporal patterns of the degree correlations between water storage variations (TWSA) and P-E anomalies. Figures 4-7 present the results of this assessment across each of the four large-scale basins forming the focus of this study. For each of the figures, time series of the spatial average TWSA and P-E anomalies are shown in order to compare their trends with the temporal behavior of the degree correlation (r_l). This comparison is helpful in determining whether the cause of trends in water storage variations (either natural or anthropogenic) influence the

analysis of hydrological consistency. For example, do the degree correlations behave differently during wet or dry periods, or when storage changes can be attributed to either natural or anthropogenic causes? In these figures, the response of the degree correlations is shown in time across the x-axis and in the spectral domain along the y-axis, for each of the three evaporation products.

4.2.1 Colorado river basin (CRB)

The Colorado River basin has experienced decadal intervals of wet and dry periods, with the start of the GRACE observing period coinciding with the end of an intense multi-year drought (Scanlon et al. 2015). During the wet period of 2004-2005 (Figures 2), the CRB region showed a corresponding increase in TWSA (Figure 4), although with a delay in time of two to three months. During this period of increase in TWSA, there was a corresponding increase in r_l (up to 0.9 for l=25 and 0.8 for l=40) until TWSA reached its peak value (November 2004 – February 2004), after which r_l decreased and showed negative values (similar but negative, i.e. -0.9 for l=25; -0.8 for l=40) during the TWSA decrease. During the dry period (i.e. 2008-2009), TWSA is correspondingly lower, but oscillating out of phase with P-E anomalies (about 2 months of lag). Similar to the wet period, there seems to be a clearer connection between the oscillation of TWSA and degree correlations in the dry period than in the rest of the study period, where there is a generally weaker connection, i.e. the variations of r_l appear random. In general, the degree correlations for small degrees have larger amplitudes than those for large degrees. This spatial disagreement in correlation might be related to the spatial and temporal distribution of runoff in the basin: a large portion of the runoff comes from snowmelt originating in the upper Colorado River Basin (Scanlon et al., 2015). Differences in absolute values and in the temporal distribution of E (especially with the MOD16 product) were evident in the degree correlation images in Figure 4, however they did not have a significant impact on the analysis in the sense of demonstrating an advantage or disadvantage over the other evaporation products in terms of hydrological consistency.

4.2.2 Niger river basin (NRB)

The TWSA in this region was characterized by an overall steady increase (5.79 mm.yr^{-1}) and a clear seasonal variability. Precipitation peaked between July and September, while TWSA peaked between September and November. Ahmed et al. (2014) attributed the increase in TWSA to an increase in precipitation in the region caused by warmer Atlantic Ocean temperatures. This trend in P was validated using multiple precipitation sources, including satellite products and rain gauges (Ahmed et al., 2014). While the GPCP data set used here did not show any increase in precipitation, neither did a recent study using rainfall estimates from the Tropical Rainfall Measuring Mission (TRMM) (Ayman and Jin, 2016), so the true cause of this trend remains somewhat unresolved. During the first two years of the GRACE observing period, the region experienced a downward trend in TWSA. During this time, the r_l values increased at the same time as TWSA decreased towards its minimum value. Then, while TWSA values were recovering, the correlation quickly decreased and became negative. This is the similar to what was observed in the CRB during the dry period. During some wet periods (e.g. July/August 2006, 2007 and 2008), when TWSA increased towards its highest value, r_l increased and was positive, but then

decreased after TWSA peaked. More generally, there seems to be a connection between r_l and the water cycle variations in this region: both high TWSA and low TWSA produced positive correlations. The transitions from positive to negative values make sense considering that when TWSA values approach zero, the observations are more uncertain, as they are affected by noise (i.e. signal to noise ratio). However, the relation might also be influenced by the lag in phase between GRACE

5 observations of TWSA and P-E. Interestingly, there also seems to be less inter-degree variability compared to other regions. This may be related to the simpler water budget in this region compared to that of the CRB and ASB, but requires further investigation.

4.2.3 Aral Sea basin (ASB)

The Aral Sea basin is an endorheic basin that has experienced a significant loss of water during the study period (-8.2 mm.yr

10 $^{-1}$), in line with the historical depletion of this inland sea in response to increased agricultural productivity (Zmijewski and Becker, 2014). Although there were short intense precipitation events during much of the study period (Figure 2), the total annual precipitation showed a negative trend of -31 mm.yr^{-1} from 2003-2008. However, water storage values increased in 2005 as a result of the construction of a dam between the north and south portions of the Aral Sea (Shi et al., 2014). During most of the study period, the r_l values oscillated in a similar way as for the CRB region: that is, a weak connection between

15 high r_l values and increasing or decreasing TWSA before reaching the local maxima or minima, respectively. Some examples of this behaviour include June–August 2008 and July-October 2009, before TWSA reaches its lowest value. In general, the r_l values decreased with increasing degree. However, inter-degree variability was more complicated in this region during several months. Although the ASB is dominantly arid, the south-east portion of the basin includes a mixture of warm and cold climates where most precipitation occurs. Due to the mismatch in resolution and/or different land cover

20 inputs, the evaporation products may represent these inter-mixed regions differently. Furthermore, glacial and snowmelt runoff present further complications to the hydrological description. These complications are reflected in the higher inter-degree variability (compared to the other regions), as well as in the slight differences in degree correlation between the evaporation products.

4.2.4 Lake Eyre basin (LEB)

25 Another endorheic basin examined here was the Lake Eyre basin, which experienced a marked increase in precipitation (Figure 2) during the rainy seasons in 2009-2011, resulting in an increase in water storage anomalies of about 40 mm/year (calculated from September 2009, to December 2011). The duration of periods in which TWSA and P-E were negatively correlated (i.e. negative r_l values) increased during this period. Total annual precipitation decreased from 2003-2006 (-23 mm.yr^{-1}), with a corresponding secular decreasing trend in TWSA of -8.26 mm.yr^{-1}. During this period however, the degree

30 correlations did not reveal any structure or indicate any connection with either P-E or TWSA. A short but intense precipitation event during the winter of 2006-2007 (Figure 2) did not seem to affect the variations in rl, relative to the earlier years. The r_l variations did show improvements however, during most of 2008 (during which precipitation was low, i.e.

P<50mm), particularly with the MOD16 evaporation product (i.e. the lowest evaporation values). In general, the r_l values generally decreased with decreasing length scales. Differences in absolute r_l values were visible between the evaporation products, but not in the overall spatial and temporal patterns. Wang et al. (2014) also studied the hydrological consistency of satellite products (TRMM P, MOD16 E and GRACE TWSA) over the LEB, as well as other predominantly arid regions of

5   the Australian continent. At the monthly scale, their study also found poor agreement between TWSA and P-E.

4.3 Applying a phase lag to GRACE data

Before GRACE can detect a significant water storage increase, the water mass from precipitation needs to build up to a significant amount within the catchment (i.e. the spatially distributed rainfall needs to accumulate in either river channels or subsurface reservoirs). This may take up to several months, as water accumulates in the soil and travels from different source

10   areas (Rieser et al., 2010). The apparent lag that GRACE data experiences relative to precipitation events has been observed in African basins (Ahmed et al., 2011; Ayman and Jin, 2016) as well as in Australia (Rieser et al., 2010; Wang et al., 2014). The clearest example of those basins studied here is in the NRB (Figure 5), where a lag of two months is evident throughout the study period. In other regions however, such as the CRB and LEB, the time needed to detect water storage changes after precipitation events tends to vary with time, perhaps due to changing spatial and temporal patterns in precipitation as well as

15   geomorphological characteristics (Ahmed et al., 2011; Wang et al., 2014).

To examine this further, a lag of one, two and three months was considered for all basins and assumed to remain constant throughout the study period. In terms of changes to the degree correlation, for the NRB it was clear that a two months lag produced an improved temporal match between the TWSA and P-E. For the other basins however, due to the changing

20   dynamics in precipitation and TWSA, a temporal match could not be satisfied at all times by using an arbitrary constant lag in GRACE. Regardless, it was found that a constant lag of two months provided a better fit compared to all alternatives (including zero lag).

Figure 8 presents a statistical summary of the mean degree correlation values over the study period comparing the original

25   analysis and using a constant lag of two months. The results are presented as boxplots, where the median is indicated as a bold black line inside a box confined by the first and third quartiles (bottom and top of the box). The whiskers below and above show a threshold of 1.5 times the inter-quartile range (IQR) below and above the first and third quartiles, defining a number of outliers outside this range. As already noted, the NRB showed a significant improvement in r_l after considering the delay, not only in terms of the median r_l value, but also in terms of the variability in the results (i.e., a smaller IQR). This

30   outcome was similar irrespective of the evaporation product used. For the CRB basin, the degree correlations did improve when using the CSIRO-PML and GLEAM products (median improved from 0.17 to 0.67, and from -0.01 to 0.64, respectively) but to a lesser extent for the MOD16 product (-0.03 to 0.29). The IQR was also reduced significantly with the CSIRO-PML product, moderately reduced with GLEAM, and did not change with the MOD16 product. The degree

correlation in the ASB region also benefited from an imposed lag in GRACE data, although there remained considerable variability in the results. The LEB region showed only a marginal increase in the amplitude of r_l and a minor reduction in the temporal variability (-0.06 to 0.14, 0.08 to 0.20 and 0.13 to 0.29 with CSIRO-PML, GLEAM and MOD16 respectively).

5 Discussion

To date, the development of methods and sensors to retrieve the various components of the water cycle has largely been undertaken independently of other (interrelated) variables, with each hydrological process presenting its own limitations, errors and retrieval challenges (McCabe et al. 2008). Therefore, irrespective of the physical constraint that the concept of hydrological consistency is based upon (i.e. mass balance), given the variability in our capacity to accurately retrieve hydrological responses via satellite-based systems, it is not unreasonable to expect that achieving hydrological consistency might be a difficult task. However, it should be reasonable to expect that in some regions, especially those where simpler and more defined water cycle behaviour dominates, that more significant and consistent agreement might be found. For this reason, the study assumed a simple water budget consisting of water storage anomalies as a function of precipitation and evaporative fluxes only and was deliberately limited to regions that would more closely reflect this simple closure assessment as much as possible. It is worth noting that an assumption of a simplified water budget in order to evaluate agreement in satellite products has been employed before. Indeed, Wang et al. (2014) applied this concept to evaluate the level of agreement between three satellite products over arid regions in Australia, assuming surface and subsurface runoff were minimal. Given the relationship between size and accuracy for GRACE data, a geographically distributed selection of basins that could fit into this simplified water budget assumption is somewhat limited. Our study consisted of four major river basins in the world, including two endorheic basins. Although having mostly an arid climate in terms of area, two of the selected basins in this study (Colorado River basin and Aral Sea basin) included regions with the presence of snow and snowmelt-dominated runoff. The inclusion of these basins was considered important in order to test the hydrological consistency concept in regions that deviated slightly from the ideal assumption.

A number of studies (Pan and Wood, 2006; Pan et al., 2008; Sheffield et al., 2009; Sahoo et al., 2011; Pan et al., 2012) have previously analysed the water budget of multiple basins using independent satellite observations (including GRACE) at the basin scale level. A distinguishing feature of the methodology employed in this study is the direct use of spherical harmonic analysis to assess the agreement between water budget components. Such an approach is useful for two reasons. The first lies within the context of processing GRACE data, which has been explored recently in Long et al. (2015). In the standard approach to using GRACE data, the scaling factors used to restore the GRACE signal after filtering (Wahr and Velicogna, 2006; Landerer and Swenson, 2012) can vary depending on the choice of LSM used to derive them. By means of spherical harmonic analysis, the effect of the filters can be incorporated directly into the other hydrological data sets (e.g. precipitation and evaporation data) without the need to choose an LSM. To our knowledge, only one other study has incorporated this

approach (Swenson 2010), with the intent of comparing two precipitation products during winter, using GRACE estimates of water storage anomalies and output from a land surface model to incorporate evaporation and runoff into the study. The second reason was to explore the potential use of the degree correlation measure (Arkani-Hamed, 1998; Tapley et al., 2004a) as a way to evaluate the hydrological consistency of satellite products. In principle, the degree correlation measure could be

5   used to incorporate the spatial dimension into the analysis by means of an approximate relation between degree (l) in spherical harmonic coefficients and spatial scales. However, the limited spatial resolution in GRACE data, further limited by the use of filters to remove errors in the data, allows for only a narrow range of length scales (500 – 800 km) to be evaluated.

The study showed that in general, the correlation between the satellite products decreased with increasing degree (i.e. smaller

10 length scales). This was expected, given the fact that GRACE errors increase with increasing degree. However, the r_l values were dominated by the temporal dimension, resulting in a distinct striping pattern. That is, the values oscillated from negative to positive values with a noisier seasonality than the underlying time series of TWSA and P-E. The fact that the amplitude in the r_l signal is high could be an indication that there is a strong temporal source of error in the combination of the satellite products. One possibility that has been explored in studies using GRACE data (Rieser et al., 2010; Ahmed et al.,

15 2011) is that water anomalies are detected some time after a precipitation event, due to the inability of GRACE to detect small-scale changes in the gravity field, and therefore the corresponding mass is detected until a sufficient amount has accumulated within the catchment via natural drainage processes.. The intensity and duration of the precipitation events, antecedent soil moisture, as well as the geomorphological characteristics of the basin would thus all influence the detection time. A simple way to account for this phenomenon was to apply a phase lag to GRACE data by a constant amount for the

20 whole study period. This seemed to improve the behaviour of degree correlation, not only in time (less variability in the results), but increased the value of r_l as well. This was particularly evident in the Niger River basin, which was expected due to the well-defined seasonal behaviour of its hydrological cycle throughout the study period, and to a lesser extent in the Colorado River basin and Aral Sea basin, where changing trends in the seasonal patterns of precipitation make it more challenging to apply this simple correction. In the Lake Eyre basin, applying a lag to GRACE data did not seem to have an

25 effect on the degree correlation. Further understanding the implications and physical rationale behind the attribution of this lag is required.

A secondary objective of this work was to determine whether differences in available evaporation products could affect the analysis i.e. was there better agreement between water storage anomalies and P-E with one particular evaporation product?

30 The products used in the study covered a wide range of resolutions (0.05°–0.25°), although the effective resolution in the analysis is determined by the truncation degree (l_{max}). Overall, results indicated that MOD16 underestimated evaporation when compared to CSIRO-PML and GLEAM, even though both the CSIRO-PML and MOD16 products are based on the Penman-Monteith equation and rely heavily on MODIS data. Several recent studies (McCabe et al., 2016; Michel et al., 2016; Miralles et al. 2016) also suggest that the MOD16 product (or variants using the PM-Mu approach) underestimate

evaporation when compared to other products (including GLEAM), and that most products show large discrepancies in reproducing results during periods of water stress. Ershadi et al. (2015) demonstrated that the parameterization of aerodynamic and surface resistances were critical controls on evaporation through both soil and vegetation. Furthermore, both GLEAM and CSIRO-PML include dynamic constraints on evaporation (stress module and soil moisture assimilation in

5   GLEAM; dynamic ratio of actual to potential evaporation at the soil surface in CSIRO-PML) that are critical in arid regions due to hydrological and plant physiological stresses and the subsequent importance of soil evaporation. In the Colorado and Aral basins, MOD16 was also lagged in phase with respect to the other evaporation products. How much these differences in phase affected the study relies in part on the ratio P/E. In the Colorado basin, it slightly affected the analysis at a few time steps (e.g. from negative to positive correlation), while in the Aral basin it did not seem to strongly affect results, as P was

10   much larger, therefore eliminating the lag from the P-E anomalies.

Apart from the limitations in evaporation modelling, the study is greatly influenced by the limitations and challenges involved in observing water storage from GRACE satellites. First, the resolution of GRACE is very limited compared to the current global satellite-based products of evaporation and precipitation. This constrains the study of hydrological consistency

15   not only in the mismatch of resolution between products, but also in that it introduces temporal error sources due to the delay in GRACE detection of surface mass changes. Because the data is smoothed to remove errors in short-scale terms (i.e. truncation of the spherical harmonic coefficients), the signal contains contamination from outside of the basins. This is a potential source of error in areas neighbouring high amplitude signals and the ocean. Furthermore, due to the lack of vertical resolution in GRACE, the effect of atmospheric fields has to be removed before producing the hydrologic product. The

20   errors in these fields ultimately contribute to errors in GRACE data (Velicogna and Wahr, 2013). One final limitation in the observational datasets used here related to the incapacity of most precipitation products to detect low intensity rainfall events (Hou et al., 2014). Determining whether or not and understanding how much these omissions affect hydrological consistency studies remains an area requiring further investigation.

25   One key motivation of the study was to assist in the validation and evaluation of individual components of the water cycle by comparing the spatial patterns between water fluxes and changes in water storage in a simple environment. Such an approach would prove particularly useful for validating evaporation products: a task that is currently challenging due to the lack of large scale spatially distributed in situ coverage. If such an approach could be first evaluated in simple water cycle systems, the potential for broader scale application in regions with more complex system behaviour would be the next logical step.

30   However, the study showed that even in these relatively simple water cycle systems, it was not possible to demonstrate a consistent hydrological agreement between current independent observations. Efforts to improve satellite-based evaporation products will continue through advances in algorithm development, increases in the observable resolution and also via the development of multi-product ensembles (with weighting based on validation analyses and uncertainty assessments) (McCabe et al., 2016; Miralles et al. 2016). The prospects for improved precipitation monitoring is also promising given the

Global Precipitation Measurement mission, which will allow for a more accurate representation of light rains - a challenge that has been a limitation in other precipitation products, including the GPCP (Huffman et al., 2001). Likewise, the next generation gravity missions (GRACE follow on and GRACE II) with the incorporation of improved sensor design (Christophe et al., 2015) are likely to provide more accurate estimates of the water storage anomalies, which are critical in

5 the determination of water budget closure studies. Efforts to evaluate the hydrological consistency should be continued in the future, as this task will help to discover the key areas of improvement in developing robust evaporation products, and potentially be used as validation tools for the next data products being developed.

6 Conclusions

Given the inherent challenges in validating satellite based products via the use of ground based observations, one of the

10 motivating elements of this study was to examine the capacity of independent observations (or estimates) of the water cycle to achieve some form of hydrological closure. To do this, the study focused on regions where it would be most expected to achieve hydrological consistency: that is, in arid regions with a simplified water budget consisting (ideally) of only precipitation and evaporation. We found that even in these simple environments, hydrological consistency was difficult to obtain throughout the study period. While there are times and locations at which some consistency was observed, there are a

15 greater number for when it is not. The lack of a persistent behaviour is problematic, both in our attempts at independently evaluating remote sensing data and also in our efforts of discriminating the uncertainty in individual products. Furthermore, although there were significant differences in evaporation estimates, especially with the MOD16 product in the Colorado and Aral basins, these differences did not play a significant role in the evaluation of hydrological consistency. Although not providing a comprehensive tool for product evaluation, the approach did help to reveal interesting spatial and temporal

20 patterns.

In general, the correlation between the satellite products was higher with smaller degrees, or larger spatial scales. In simple water cycle systems such as in the Niger River basin, the correlation followed cyclical patterns along with the water storage anomalies i.e. it increased along with water storage anomalies up to the point where these peaked, then decreased along with

25 them up to the point where these were minimal, then the same pattern repeated but with negative anomalies. This indicates that, at the least, the correlations are not random, but roughly follow the cyclical variations within the basin. It is also reasonable to expect low agreement when fluxes and/or water storage anomalies are minimal, explaining some of the cyclical nature in the correlation. A lag between GRACE and precipitation data was also considered, and it was shown that imposing even a simple correction (i.e. a constant phase shift to GRACE data) greatly improved the agreement, both in

30 average degree correlation and variability of the results in time.

The lack of persistent agreement in other regions may be explained in part by the added complexities that limit the validity of the assumption of a simple water cycle, such as snow melt runoff, complex geomorphology, changing patterns in precipitation as well as anthropogenic influence on the water system. Further limitations to hydrological consistency include the many challenges that still exist in the large-scale retrieval of precipitation, evaporation and GRACE data. As the

5   algorithms and input data required for the estimation of the water cycle components from satellites improve, it will be important to further explore both water budget closure studies to evaluate the agreement of the observations at the basin scale level and hydrological consistency approaches as a potential tool in the evaluation of new data products.

Acknowledgments. Research reported in this publication was supported by the King Abdullah University of Science and
10   Technology (KAUST).

References

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P.-P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., Gruber, A., Susskind, J., Arkin, P. and Nelkin, E.: The version-2 Global Precipitation Climatology Project (GPCP) monthly precipitation analysis (1979-present), Journal of Hydrometeorology, 4(6), 1147-1167, doi:10.1175/1525-
15   7541(2003)004<1147:TVGPCP>2.0.CO;2, 2003.

Ahmed, M., Sultan, M., Wahr, J., Yan, E., Milewski, A., Sauck, W., Becker, R. and Welton, B.: Integration of GRACE (Gravity Recovery and Climate Experiment) data with traditional data sets for a better understanding of the time-dependent water partitioning in african watersheds, Geology, doi:10.1130/G31812.1, 2011.

Ahmed, M., Sultan, M., Wahr, J. and Yan, E.: The use of GRACE data to monitor natural and anthropogenic induced
20   variations in water availability across Africa, Earth-Science Reviews, 136, 289-300, doi:http://dx.doi.org/10.1016/j.earscirev.2014.05.009, 2014.

Arkani-Hamed, J.: The lunar mascons revisited, Journal of Geophysical Research: Planets, 103(E2), 3709-3739, doi:10.1029/97JE02815, 1998.

Ayman, H. and Shuanggen, J.: Water storage changes and balances in Africa observed by GRACE and hydrologic models,
25   Geodesy and Geodynamics, doi:http://dx.doi.org/10.1016/j.geog.2016.03.002, 2016.

Cheng, M., Ries, J. C. and Tapley, B. D.: Variations of the earth's figure axis from satellite laser ranging and GRACE, Journal of Geophysical Research: Solid Earth, 116(B1), doi:10.1029/2010JB000850, 2011.

Cheng, M., Tapley, B. D. and Ries, J. C.: Deceleration in the earth's oblateness, Journal of Geophysical Research: Solid Earth, 118(2), 740-747, doi:10.1002/jgrb.50058, 2013.

30   Christophe, B., Boulanger, D., Foulon, B., Huynh, P.-A., Lebat, V., Liorzou, F. and Perrot, E.: A new generation of ultra-sensitive electrostatic accelerometers for {gRACE} follow-on and towards the next generation gravity missions, Acta Astronautica, 117, 1-7, doi:http://dx.doi.org/10.1016/j.actaastro.2015.06.021, 2015.

Cleugh, H. A., Leuning, R., Mu, Q. and Running, S. W.: Regional evaporation estimates from flux tower and MODIS satellite data, Remote Sensing of Environment, 106(3), 285-304, doi:http://dx.doi.org/10.1016/j.rse.2006.07.007, 2007.

Crow, W. T.: A novel method for quantifying value in spaceborne soil moisture retrievals, Journal of Hydrometeorology, 8(1), 56-67, doi:10.1175/JHM553.1, 2007.

5   Duan, X., Guo, J., Shum, C. and Wal, W. van der: On the postprocessing removal of correlated errors in GRACE temporal gravity field solutions, Journal of Geodesy, 83(11), 1095-1106, doi:10.1007/s00190-009-0327-0, 2009.

Ershadi, A., McCabe, M., Evans, J., Chaney, N. and Wood, E.: Multi-site evaluation of terrestrial evaporation models using FLUXNET data, Agricultural and Forest Meteorology, 187(0), 46-61, doi:http://dx.doi.org/10.1016/j.agrformet.2013.11.008, 2014.

10   Ershadi, A., McCabe, M., Evans, J. and Wood, E.: Impact of model structure and parameterization on penman-Monteith type evaporation models, Journal of Hydrology, 525, 521-535, doi:http://dx.doi.org/10.1016/j.jhydrol.2015.04.008, 2015.

Famiglietti, J. S., Lo, M., Ho, S. L., Bethune, J., Anderson, K. J., Syed, T. H., Swenson, S. C., Linage, C. R. de and Rodell, M.: Satellites measure recent rates of groundwater depletion in California's central valley, Geophysical Research Letters, 38(3), doi:10.1029/2010GL046442, 2011.

15   Famiglietti, J. S., Cazenave, A., Eicker, A., Reager, J. T., Rodell, M. and Velicogna, I.: Satellites provide the big picture, Science, 349(6249), 684–685, doi:10.1126/science.aac9238, 2015.

Fisher, J. B., Tu, K. P. and Baldocchi, D. D.: Global estimates of the land-atmosphere water flux based on monthly AVHRR and iSLSCP-iI data, validated at 16 FLUXNET sites, Remote Sensing of Environment, 112(3), 901-919, doi:http://dx.doi.org/10.1016/j.rse.2007.06.025, 2008.

20   Gash, J. H. C.: An analytical model of rainfall interception by forests, Quarterly Journal of the Royal Meteorological Society, 105(443), 43-55, doi:10.1002/qj.49710544304, 1979.

Hou, A. Y., Kakar, R. K., Neeck, S., Azarbarzin, A. A., Kummerow, C. D., Kojima, M., Oki, R., Nakamura, K. and Iguchi, T.: The global precipitation measurement mission, Bulletin of the American Meteorological Society, 95(5), 701-722, doi:10.1175/BAMS-D-13-00164.1, 2014.

25   Huffman, G. J., Adler, R. F., Arkin, P., Chang, A., Ferraro, R., Gruber, A., Janowiak, J., McNab, A., Rudolf, B. and Schneider, U.: The global precipitation climatology project (gPCP) combined precipitation dataset, Bulletin of the American Meteorological Society, 78(1), 5-20, doi:10.1175/1520-0477(1997)078<0005:TGPCPG>2.0.CO;2, 1997.

Huffman, G. J., Adler, R. F., Morrissey, M. M., Bolvin, D. T., Curtis, S., Joyce, R., McGavock, B. and Susskind, J.: Global precipitation at one-degree daily resolution from multisatellite observations, Journal of Hydrometeorology, 2(1), 36-50,
30   doi:10.1175/1525-7541(2001)002<0036:GPAODD>2.0.CO;2, 2001.

Jimenez, C., Prigent, C., Mueller, B., Seneviratne, S. I., McCabe, M. F., Wood, E. F., Rossow, W. B., Balsamo, G., Betts, A. K., Dirmeyer, P. A., Fisher, J. B., Jung, M., Kanamitsu, M., Reichle, R. H., Reichstein, M., Rodell, M., Sheffield, J., Tu, K. and Wang, K.: Global intercomparison of 12 land surface heat flux estimates, Journal of Geophysical Research: Atmospheres, 116(D2), doi:10.1029/2010JD014545, 2011.

Kalma, J. D., McVicar, T. R. and McCabe, M. F.: Estimating land surface evaporation: A review of methods using remotely sensed surface temperature data, Surveys in Geophysics, 29(4), 421-469, doi:10.1007/s10712-008-9037-z, 2008.

Kottek, M., Grieser, J., Beck, C., Rudolf, B. and Rubel, F.: World map of the Köppen-geiger climate classification updated, Meteorologische Zeitschrift, 15(3), 259-263, doi:doi:10.1127/0941-2948/2006/0130, 2006.

5   Landerer, F. W. and Swenson, S. C.: Accuracy of scaled GRACE terrestrial water storage estimates, Water Resources Research, 48(4), doi:10.1029/2011WR011453, 2012.

Leuning, R., Zhang, Y. Q., Rajaud, A., Cleugh, H. and Tu, K.: A simple surface conductance model to estimate regional evaporation using mODIS leaf area index and the penman-monteith equation, Water Resources Research, 44(10), doi:10.1029/2007WR006562, 2008.

10   Liu, Y., Dorigo, W., Parinussa, R., Jeu, R. de, Wagner, W., McCabe, M., Evans, J. and Dijk, A. van: Trend-preserving blending of passive and active microwave soil moisture retrievals, Remote Sensing of Environment, 123(0), 280-297, doi:http://dx.doi.org/10.1016/j.rse.2012.03.014, 2012.

Long, D., Longuevergne, L. and Scanlon, B. R.: Global analysis of approaches for deriving total water storage changes from GRACE satellites, Water Resources Research, doi:10.1002/2014WR016853, 2015.

15   Martinec, Z.: Program to calculate the spectral harmonic expansion coefficients of the two scalar fields product, Computer Physics Communications, 54(1), 177-182, doi:http://dx.doi.org/10.1016/0010-4655(89)90043-X, 1989.

McCabe, M. F. and Wood, E. F.: Scale influences on the remote estimation of evapotranspiration using multiple satellite sensors, Remote Sensing of Environment, 105(4), 271-285, doi:http://dx.doi.org/10.1016/j.rse.2006.07.006, 2006.

McCabe, M., Wood, E., Wójcik, R., Pan, M., Sheffield, J., Gao, H. and Su, H.: Hydrological consistency using multi-sensor

20   remote sensing data for water and energy cycle studies, Remote Sensing of Environment, 112(2), 430-444, doi:http://dx.doi.org/10.1016/j.rse.2007.03.027, 2008.

McCabe, M. F., Ershadi, A., Jimenez, C., Miralles, D. G., Michel, D. and Wood, E. F.: The GEWEX LandFlux project: Evaluation of model evaporation using tower-based and globally gridded forcing data, Geoscientific Model Development, 9(1), 283-305, doi:10.5194/gmd-9-283-2016, 2016.

25   Michel, D., Jiménez, C., Miralles, D. G., Jung, M., Hirschi, M., Ershadi, A., Martens, B., McCabe, M. F., Fisher, J. B., Mu, Q., Seneviratne, S. I., Wood, E. F. and Fernández-Prieto, D.: The WACMOS-ET project - Part 1: Tower-scale evaluation of four remote-sensing-based evapotranspiration algorithms, Hydrology and Earth System Sciences, 20(2), 803-822, doi:10.5194/hess-20-803-2016, 2016.

Milewski, A., Sultan, M., Jayaprakash, S. M., Balekai, R. and Becker, R.: RESDEM, a tool for integrating temporal remote

30   sensing data for use in hydrogeologic investigations, Computers & Geosciences, 35(10), 2001-2010, doi:http://dx.doi.org/10.1016/j.cageo.2009.02.010, 2009.

Miralles, D. G., Holmes, T. R. H., De Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A. and Dolman, A. J.: Global land-surface evaporation estimated from satellite-based observations, Hydrology and Earth System Sciences, 15(2), 453-469, doi:10.5194/hess-15-453-2011, 2011.

Miralles, D. G., Jiménez, C., Jung, M., Michel, D., Ershadi, A., McCabe, M. F., Hirschi, M., Martens, B., Dolman, A. J., Fisher, J. B., Mu, Q., Seneviratne, S. I., Wood, E. F. and Fernández-Prieto, D.: The WACMOS-ET project - Part 2: Evaluation of global terrestrial evaporation data sets, Hydrology and Earth System Sciences, 20(2), 823-842, doi:10.5194/hess-20-823-2016, 2016.

5   Mu, Q., Heinsch, F. A., Zhao, M. and Running, S. W.: Development of a global evapotranspiration algorithm based on MODIS and global meteorology data, Remote Sensing of Environment, 111(4), 519-536, doi:http://dx.doi.org/10.1016/j.rse.2007.04.015, 2007.

Mu, Q., Zhao, M. and Running, S. W.: Improvements to a MODIS global terrestrial evapotranspiration algorithm, Remote Sensing of Environment, 115(8), 1781-1800, doi:http://dx.doi.org/10.1016/j.rse.2011.02.019, 2011.

10   Pan, M. and Wood, E. F.: Data assimilation for estimating the terrestrial water budget using a constrained Ensemble Kalman Filter, Journal of Hydrometeorology, 7(3), 534-547, doi:10.1175/JHM495.1, 2006.

Pan, M., Wood, E. F., Wójcik, R. and McCabe, M. F.: Estimation of regional terrestrial water cycle using multi-sensor remote sensing observations and data assimilation, Remote Sensing of Environment, 112(4), 1282-1294, doi:http://dx.doi.org/10.1016/j.rse.2007.02.039, 2008.

15   Pan, M., Sahoo, A. K., Troy, T. J., Vinukollu, R. K., Sheffield, J. and Wood, E. F.: Multisource estimation of long-term terrestrial water budget for major global river basins, Journal of Climate, 25(9), 3191-3206, doi:10.1175/JCLI-D-11-00300.1, 2012.

Rieser, D., Kuhn, M., Pail, R., Anjasmara, I. M. and Awange, J.: Relation between GRACE-derived surface mass variations and precipitation over Australia, Australian Journal of Earth Sciences, 57(7), 887-900, doi:10.1080/08120099.2010.512645,

20   2010.

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin*, J. K., Walker, J. P., Lohmann, D. and Toll, D.: The global land data assimilation system, Bulletin of the American Meteorological Society, 85(3), 381-394, doi:10.1175/BAMS-85-3-381, 2004.

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich,

25   J., Bosilovich, M., Entin*, J. K., Walker, J. P., Lohmann, D. and Toll, D.: The global land data assimilation system, Bulletin of the American Meteorological Society, 85(3), 381-394, doi:10.1175/BAMS-85-3-381, 2004.

Rodell, M., Velicogna, I. and Famiglietti, J. S.: Satellite-based estimates of groundwater depletion in india, Nature, 460(7258), 999-1002, doi:10.1038/nature08238, 2009.

Sahoo, A. K., Pan, M., Troy, T. J., Vinukollu, R. K., Sheffield, J. and Wood, E. F.: Reconciling the global terrestrial water

30   budget using satellite remote sensing, Remote Sensing of Environment, 115(8), 1850-1865, doi:http://dx.doi.org/10.1016/j.rse.2011.03.009, 2011.

Scanlon, B. R., Zhang, Z., Reedy, R. C., Pool, D. R., Save, H., Long, D., Chen, J., Wolock, D. M., Conway, B. D. and Winester, D.: Hydrologic implications of GRACE satellite data in the colorado river basin, Water Resources Research, 51(12), 9891–9903, doi:10.1002/2015WR018090, 2015.

Sheffield, J., Goteti, G. and Wood, E. F.: Development of a 50-year high-resolution global dataset of meteorological forcings for land surface modeling, Journal of Climate, 19(13), 3088-3111, doi:10.1175/JCLI3790.1, 2006.

Sheffield, J., Ferguson, C. R., Troy, T. J., Wood, E. F. and McCabe, M. F.: Closing the terrestrial water budget from satellite remote sensing, Geophysical Research Letters, 36(7), doi:10.1029/2009GL037338, 2009.

5 Shi, W., Wang, M. and Guo, W.: Long-term hydrological changes of the Aral Sea observed by satellites, Journal of Geophysical Research: Oceans, 119(6), 3313-3326, doi:10.1002/2014JC009988, 2014.

Sun, A. Y.: Predicting groundwater level changes using GRACE data, Water Resources Research, 49(9), 5900-5912, doi:10.1002/wrcr.20421, 2013.

Swenson, S.: Assessing high-latitude winter precipitation from global precipitation analyses using GRACE, Journal of 10 Hydrometeorology, 11(2), 405-420, doi:10.1175/2009JHM1194.1, 2010.

Swenson, S. and Wahr, J.: Methods for inferring regional surface-mass anomalies from gravity recovery and climate experiment (GRACE) measurements of time-variable gravity, Journal of Geophysical Research: Solid Earth, 107(B9), ETG 3-1-ETG3-13, doi:10.1029/2001JB000576, 2002.

Swenson, S. and Wahr, J.: Post-processing removal of correlated errors in GRACE data, Geophysical Research Letters, 15 33(8), doi:10.1029/2005GL025285, 2006.

Swenson, S., Famiglietti, J., Basara, J. and Wahr, J.: Estimating profile soil moisture and groundwater variations using GRACE and oklahoma mesonet soil moisture data, Water Resources Research, 44(1), doi:10.1029/2007WR006057, 2008.

Tapley, B. D., Bettadpur, S., Ries, J. C., Thompson, P. F. and Watkins, M. M.: GRACE measurements of mass variability in the earth system, Science, 305(5683), 503-505, doi:10.1126/science.1099192, 2004a.

20 Tapley, B. D., Bettadpur, S., Watkins, M. and Reigber, C.: The gravity recovery and climate experiment: Mission overview and early results, Geophysical Research Letters, 31(9), doi:10.1029/2004GL019920, 2004b.

Voss, K. A., Famiglietti, J. S., Lo, M., Linage, C. de, Rodell, M. and Swenson, S. C.: Groundwater depletion in the middle east from gRACE with implications for transboundary water management in the tigris-euphrates-western iran region, Water Resources Research, 49(2), 904-914, doi:10.1002/wrcr.20078, 2013.

25 Wahr, J., Swenson, S. and Velicogna, I.: Accuracy of GRACE mass estimates, Geophysical Research Letters, 33(6), doi:10.1029/2005GL025305, 2006.

Wang, H., Wu, P. and Wang, Z.: An approach for spherical harmonic analysis of non-smooth data, Computers & Geosciences, 32(10), 1654-1668, doi:http://dx.doi.org/10.1016/j.cageo.2006.03.004, 2006.

Wang, H., Guan, H., Gutiérrez-Jurado, H. A. and Simmons, C. T.: Examination of water budget using satellite products over 30 Australia, Journal of Hydrology, 511, 546-554, doi:http://dx.doi.org/10.1016/j.jhydrol.2014.01.076, 2014.

Wang, K. and Dickinson, R. E.: A review of global terrestrial evapotranspiration: Observation, modeling, climatology, and climatic variability, Reviews of Geophysics, 50(2), doi:10.1029/2011RG000373, 2012.

Zhang, Y., Leuning, R., Hutley, L. B., Beringer, J., McHugh, I. and Walker, J. P.: Using long-term water balances to parameterize surface conductances and calculate evaporation at 0.05° spatial resolution, Water Resources Research, 46(5), doi:10.1029/2009WR008716, 2010.

Zhang, Y., Leuning, R., Chiew, F. H. S., Wang, E., Zhang, L., Liu, C., Sun, F., Peel, M. C., Shen, Y. and Jung, M.: Decadal

5   trends in evaporation from global energy and water balances, Journal of Hydrometeorology, 13(1), 379-391, doi:10.1175/JHM-D-11-012.1, 2012.

Zmijewski, K. and Becker, R.: Estimating the effects of anthropogenic modification on water balance in the aral sea watershed using GRACE: 2003-12, Earth Interactions, 18(3), 1-16, doi:10.1175/2013EI000537.1, 2014.

Product name	Spatial resolution	Time span	Reference
Evaporation			
MOD16 (A2)	0.05°	2000 - 2013	Mu et al. (2011)
CSIRO-PML	0.25°	1981 - 2011	Zhang et al. (2012)
GLEAM (v2A)	0.25°	1980 - 2011	Miralles et al. (2011)
Water storage			
GRACE (CSR RL05 GSM)	333 km (l_{max} = 60)	2003 - present	Tapley et al. (2004)
Precipitation			
GPCP (1DD v1.2)	1°	1996 - present	Huffman et al. (2001)

Table 1. Description of the satellite products used in this study. The temporal resolution is daily except for MOD16 (8-daily) and GRACE (monthly).
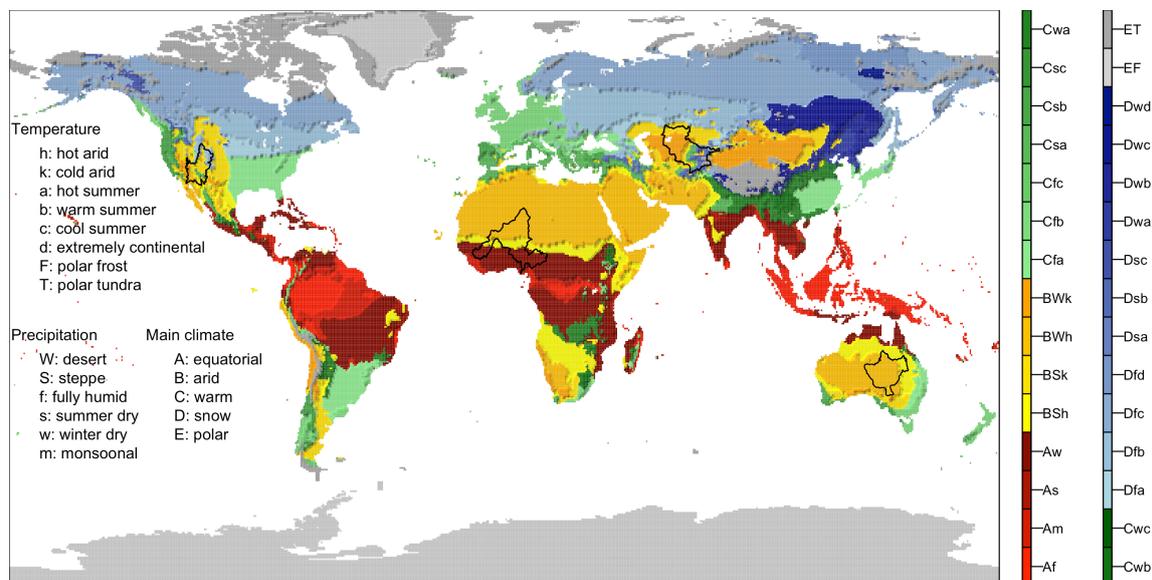
Figure 1: Selected study basins used within the analysis. Criteria for the selection of basins included: predominantly arid
climate (more than 50% areal coverage with any of the arid Köppen climates: BWk, BWh, BSk or BSh), size, geographical
location and amplitude and trends in the water storage variations.
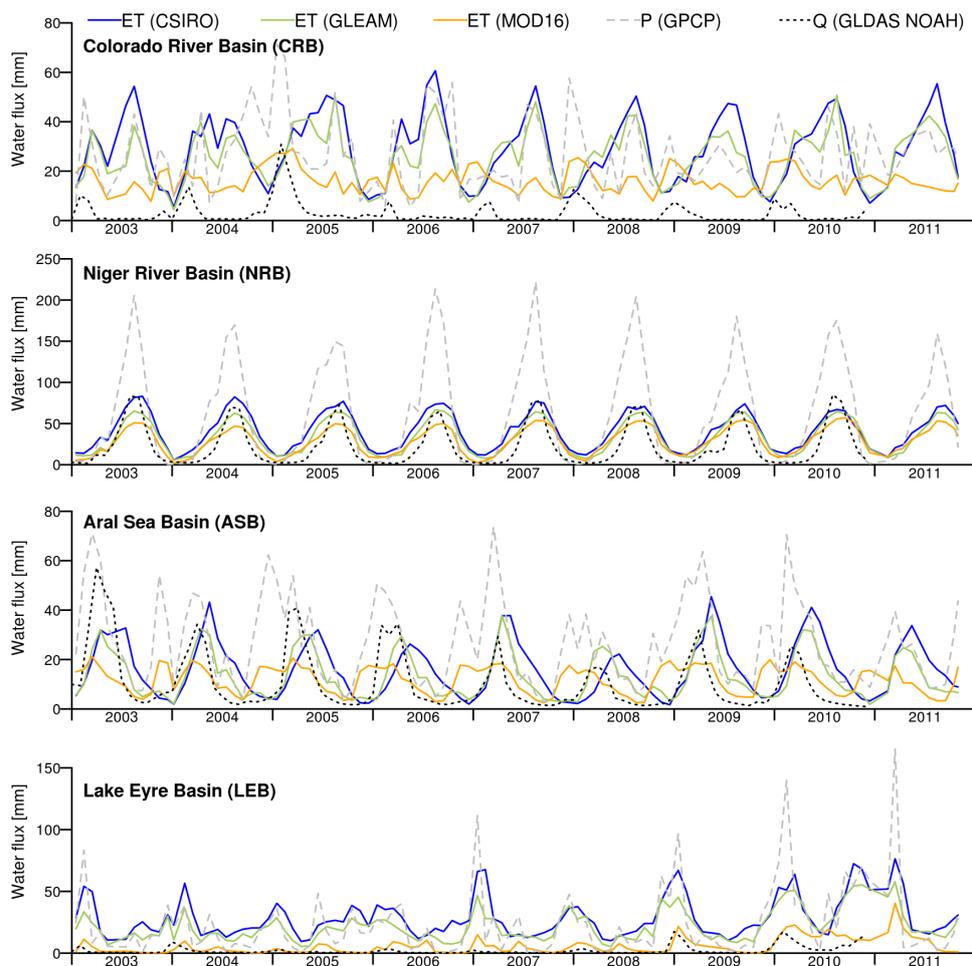
Figure 2. Average P, E and Q fluxes within the four study basins for the period 2003-2011. Three evaporation (E) data sets

5   were used in this study, including CSIRO-PML, MOD16 and GLEAM (see Table 1 for details).
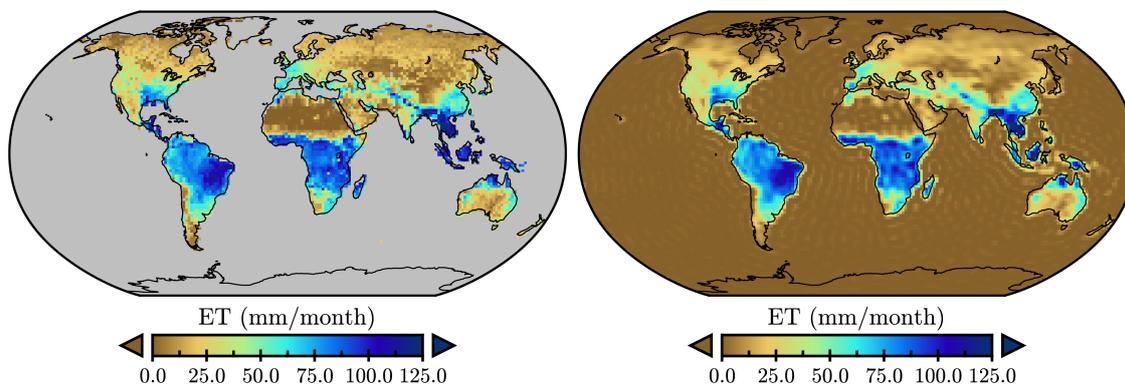
Figure 3. Left: CSIRO-PML monthly evaporation (E) for April 2003. Right: the same data set after spherical harmonic analysis and synthesis of the evaporation data. Missing data are set to zero. The data appears smoothed because it is an approximation (equation 1) limited by l_{max}. There are also other effects such as ringing that are visible in regions where the data is close to zero.
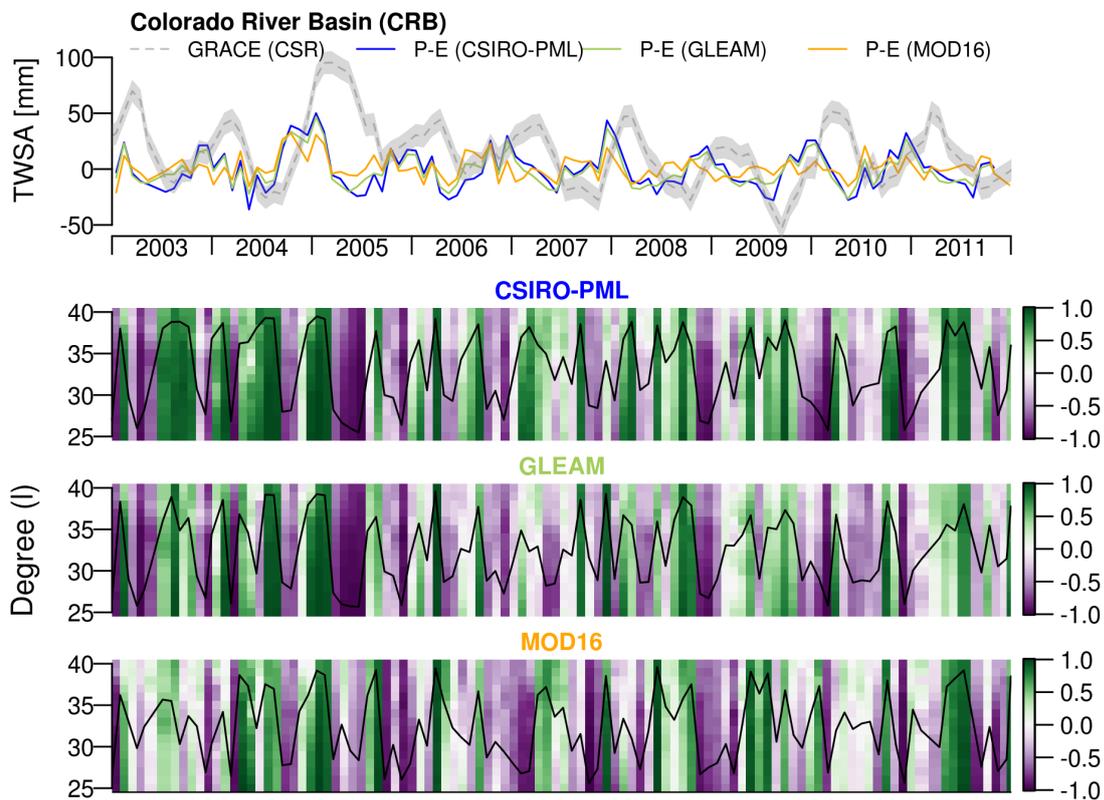
Figure 4. Top: anomalies of the terrestrial water storage (TWSA) observed by GRACE (with 20 mm uncertainty bounds) and
P-E using three global evaporation products over the Colorado river basin. Below: varying degree correlation measure (r_l)
with time and degree (from 25 to 40) using the three global evaporation products. The average r_l is shown as a time series
5   (black line). The degree correlation measure can range from -1 to 1 as shown in the color scale on the secondary axis.
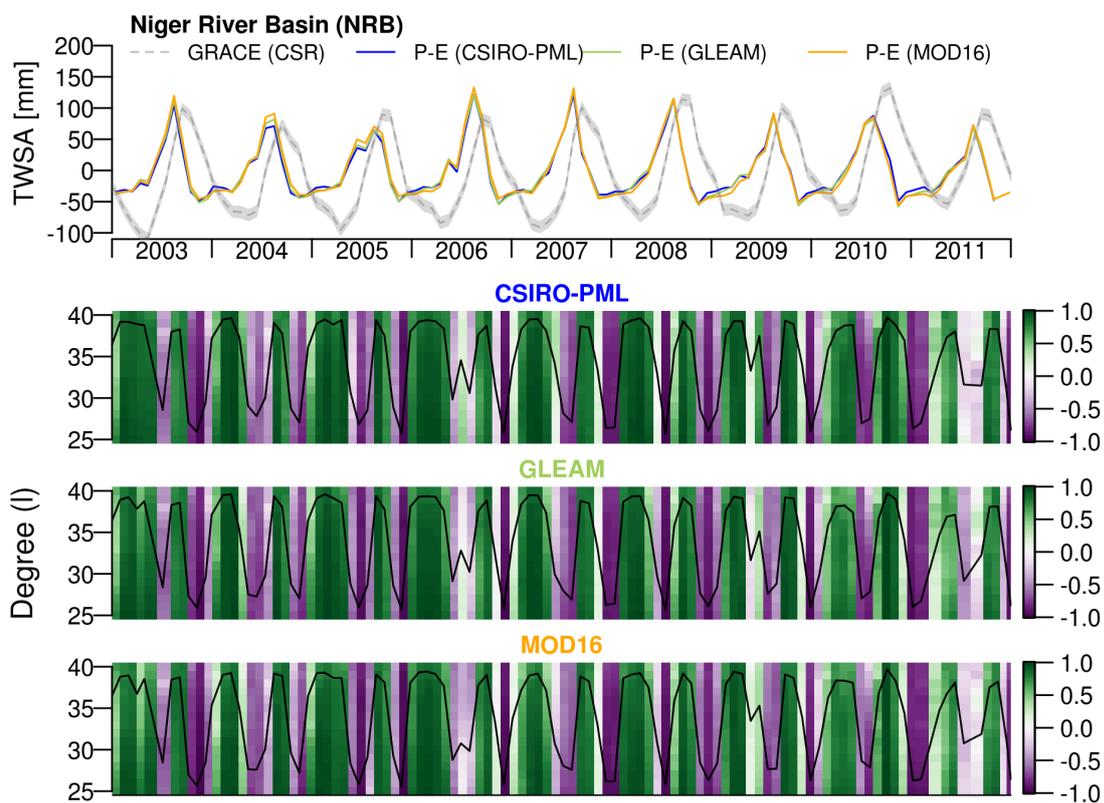
Figure 5. Top: anomalies of the terrestrial water storage (TWSA) observed by GRACE (with 20 mm uncertainty bounds) and P-E using three global evaporation products over the Niger river basin. Below: varying degree correlation measure (r_l) with time and degree (from 25 to 40) using the three global evaporation products. The average r_l is shown as a time series (black line). The degree correlation measure can range from -1 to 1 as shown in the color scale on the secondary axis.
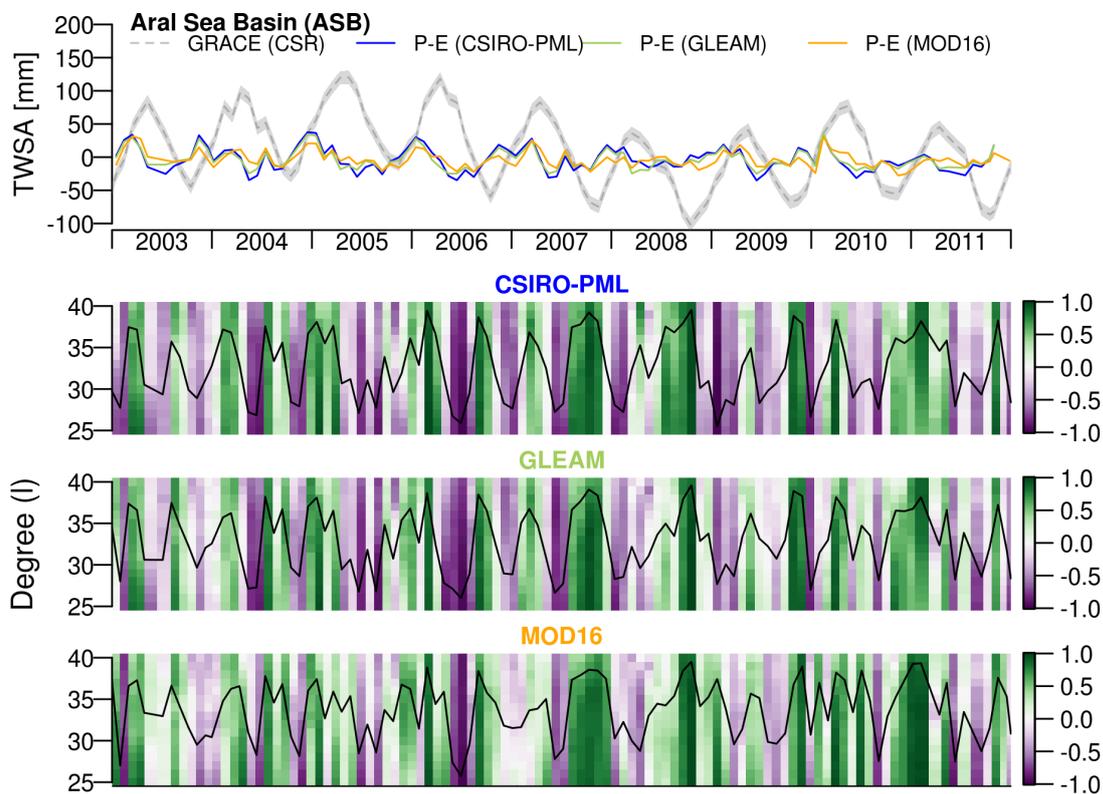
Figure 6. Top: anomalies of the terrestrial water storage (TWSA) observed by GRACE (with 20 mm uncertainty bounds) and P-E using three global evaporation products over the Aral basin. Below: varying degree correlation measure (r_l) with time and degree (from 25 to 40) using the three global evaporation products. The average r_l is shown as a time series (black line).

5   The degree correlation measure can range from -1 to 1 as shown in the color scale on the secondary axis.
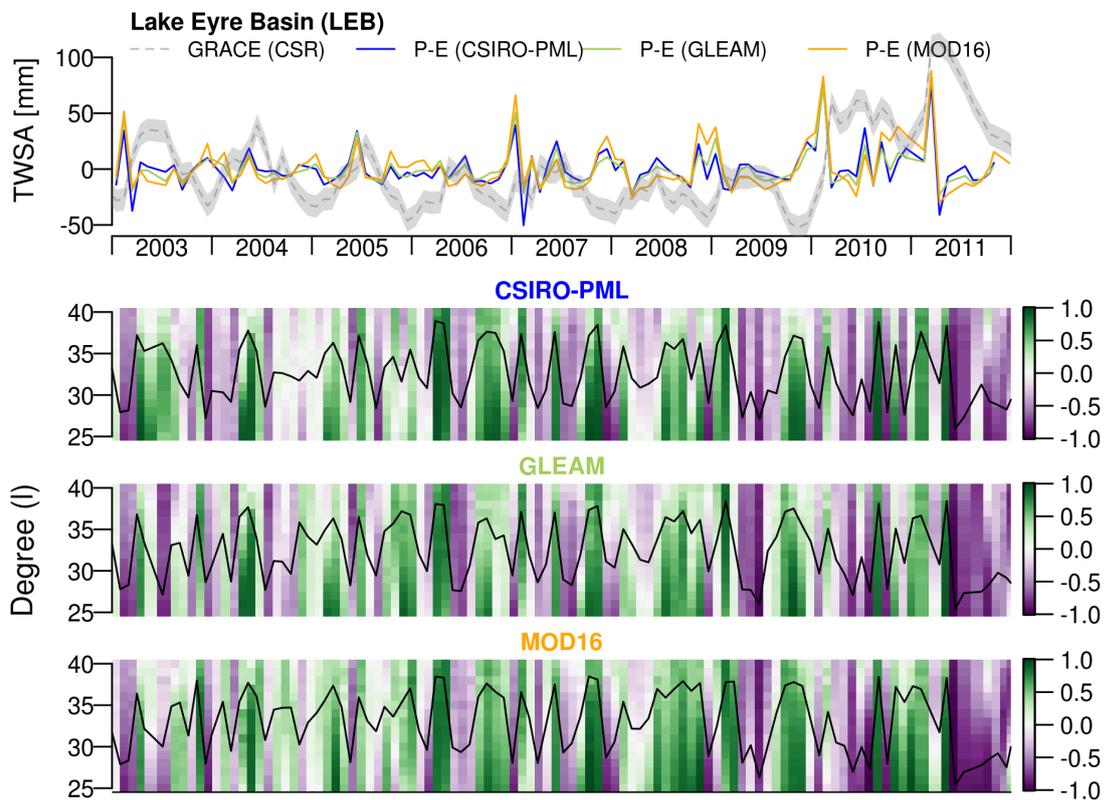
Figure 7. Top: anomalies of the terrestrial water storage (TWSA) observed by GRACE (with 20 mm uncertainty bounds) and P-E using three global evaporation products over the Lake Eyre basin. Below: varying degree correlation measure (r_l) with time and degree (from 25 to 40) using the three global evaporation products. The average r_l is shown as a time series (black line). The degree correlation measure can range from -1 to 1 as shown in the color scale on the secondary axis.
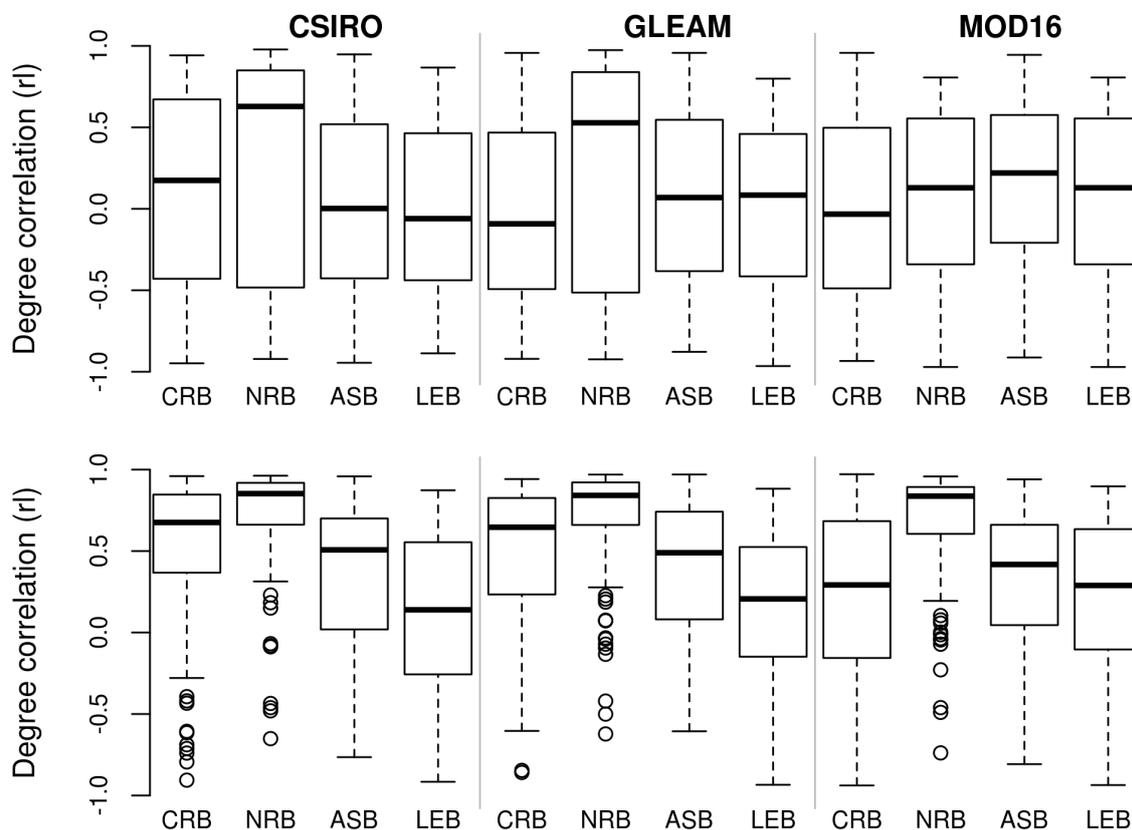
Figure 8. Top: average degree correlation statistics per study region and evaporation product. Bottom: GRACE data were shifted by two months to match the phase with P-E anomalies. The boxplots show the first, second (median) and third quartiles. Outliers, defined as data outside the 1.5 inter-quartile range (IQR) whiskers below or above the first and third quartiles are shown as circles.