Comments to the authors by Prof. Dr. Marc Bierkens

This paper seeks to evaluate the consistency between independently obtained remotely sensed water cycle products and at a comparable spatio-temporal scale. However, the focus is on Evap as 3 Evap products are used and one GRACE and one Precip product. The novelty of this study is to look at consistency instead of uncertainty and/or water balance closure. This by itself is worthy for publication. However, there are a number of issues that need to be resolved as brought forward by the two reviewers. The most important issues are with the structure and readibility of the paper which could be much improved. Most significantly:

1) The term "hydrological consistency" should be better defined.

2) The authors should more clearly state the objective(s) of the paper, state it only in one paragraph and refer to that later in the discussion of the results.

In their rebuttal the authors already provide a much clearer description of their purpose:

"Ideally, these observations should be hydro- logically consistent: that is, an observed rainfall event should cause a corresponding change in soil moisture, for instance. Likewise, a reduction in soil moisture should be reflected by an increased flux of evaporation. Consistency is just another term that en- compasses the expectation of a water budget: changes in one term should be reflected in others. While this has been explored qualitatively in the past (McCabe et al. 2008), here we wanted to determine if the method (using spherical harmonics) could reveal some level of agreement between spatial (and temporal) patterns of these independent hydrological variables."

After that they could then state the objective as" So the objective is to check this hydrologic consistency for a number of basins where hydrological processes are relatively well defined."

3) State clearly in the introduction what the novel features are of this work (checking consistency instead of water balance closure and/or uncertainty and doing this at a comparable spatio-temporal scale).

4) More clearly define what you mean by a "simple hydrological system" or "relatively well defined hydrological processes". What conditions have been used to select the basins.

5) Consider changing the title somewhat. As the focus is on Evap it could for instance be:

Evaluating the hydrological consistency among satellite based water cycle components with a focus on evaporation products

Author’s response to the Editor’s comments

Dear Prof. Dr. Marc Bierkens,

Thank you for your positive comments. In revising our manuscript, we have addressed all of the thoughtful and insightful comments put forward by the reviewers, as well as responded to the five major points you have detailed in your correspondence.

Following your suggestion, we more clearly state that the focus of the paper is to evaluate evaporation products. To reflect this, the title of the paper has been changed to: “Evaluating the hydrological consistency of evaporation products using satellite-based gravity and rainfall data”

Changes have also been reflected in a revised Abstract, Introduction, and Discussion sections, which have been subject to considerable and focused modification. Some of these are summarized below, followed directly by a complete list of changes.

- In the Introduction Section, the manuscript now addresses the challenges and issues associated with assessing large-scale evaporation products, and we have included further details as a result of the discussion with the reviewers. The definition of hydrological consistency is now more clearly articulated. The objective of the study, which is related to employing this concept of hydrological consistency, is restated in the revised Introduction.

- Another issue that was highlighted in your response was to clearly state the conditions that have been used to select basins with relatively simple interactions. We have made sure to include these early in the manuscript. In addition, we have more clearly highlighted the novelties of the study, which include (in part) the approach of hydrological consistency to evaluate evaporation products, as well as the spherical harmonics degree correlation approach to intercompare the products.

- Related to this last point, we have responded to the reviewer’s request for a more detailed explanation of the spherical harmonics concept. These include:

    - From the Methodology Section, Lines 20 in page 7 through line 11, page 8, which introduce some of the theory pertaining to GRACE and the use of spherical harmonics, have been moved up to the Data Section under the subsection “GRACE water storage anomalies”.

    - The introduction now contains a brief, but necessary introduction to this concept (to further highlight the novelty aspects of the paper).

    - In addition, lines 33 (page 2) through line 9 (page 3) have been moved to the Data section under the subsection “GRACE water storage anomalies”

- A description of the evaporation (and precipitation) data sets was also revised and now includes information on the inputs requirements, as well as a more detailed description. In this section, we have also made sure to better explain the selection process for the study basins. Some additional details on the selected basins that were previously included in the Results were moved to the Data section.

- In the Results Section, we have added some comments on the interpretation of the lag phase found in GRACE data. This includes new information pertaining to the base flow index (BFI; Beck et al., 2013), as suggested by one of the reviewers.

- The Discussion Section was also extensively revised. We have now divided this Section under three sub-headings: (1) Challenges to implementing hydrological consistency, (2) Temporal lag in terrestrial storage response, and (3) Discriminating between satellite evaporation products. We believe this will allow the reader to more clearly identify and appreciate the discussion generated from our study. We have also added a supplementary Figure and reference to it in the Discussion Section in order to address the use of another precipitation product in the study, as suggested by both reviewers.

- Finally, the abstract was also extensively revised based on the comments by the reviewers. The abstract has now a briefer introduction, and focuses more on the novelty and results of the study.

We believe that this newly revised manuscript better presents the findings of our study. We thank you for time and consideration.

List of changes to the manuscript

Abstract

Page 1, lines 7-9. Modified text:

Advances in ~~multi-satellite~~space-~~-~~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~~develop regional- -to -global-~~-~~scale estimates of evaporation, ~~products have been developed,~~ offering insights into ~~the~~ this key component of the hydrological cycle.~~multi-scale behaviour and variability of hydrological states and fluxes.~~

Page 1, lines 9-22 Removed completely, inserted:

> However, the evaluation of large-scale evaporation products is not a straightforward task. While a number of studies have intercompared a range of these products by examining the variance amongst them or by comparison of pixel-scale retrievals against ground-based observations, there is a need to explore more appropriate techniques to comprehensively evaluate flux estimates. One possible approach is to establish the level of product agreement between related hydrological components: for instance, how well do evaporation patterns and response match with precipitation or water storage changes. To assess the suitability of this "consistency"-based approach for evaluating evaporation products, we identified four globally distributed basins in arid- and semi-arid environments, including the Colorado River basin, the Niger River basin, the Aral Sea basin and the Lake Eyre basin. In an effort to identify product quality, three satellite-based global evaporation products including CSIRO-PML, MOD16 and GLEAM were evaluated against rainfall data from GPCP along with GRACE water storage anomalies. To ensure a fair comparison, we evaluated consistency using a degree correlation approach after transforming both evaporation and precipitation data into spherical harmonics.

Page 1, lines 22-24

Overall we found~~, 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 dryland environments.

Page 1, lines 24-25 Inserted:

> Indeed, the degree correlation showed oscillating values between periods of low and high water storage changes, with a phase difference of about 2-3 months. Interestingly, after imposing a simple lag in GRACE data to account for delayed surface runoff or baseflow components, there was an improved match in terms of degree correlation in the Niger River basin. Significant improvements to the degree correlations (from ~0 to about 0.6) were also found in the Colorado River basin for both the CSIRO-PML and GLEAM products (MOD16 showed only half of that

improvement). In other basins, the variability in the temporal pattern of degree correlations was still considerable and hindered any differentiation between the evaporation products. Even so, it was found that a constant lag of two months provided a better fit compared to other alternatives, including a zero lag. Regardless of this finding, from a product assessment perspective, it was concluded that no significant or persistent advantage could be identified across any of the three evaporation products in terms of a sustained hydrological consistency with precipitation and water storage anomaly data. The results of this analysis have implications in terms of the confidence that can be placed in independent retrievals of the hydrological cycle, raises questions on inter-product quality and highlights the need for additional techniques to evaluate large-scale products.

Introduction

Page 1, line 27 Inserted:

Space-based observations of the Earth system have provided the capacity to retrieve information across a wide-range of land surface hydrological components and an opportunity to characterize terrestrial processes in space and time. Indeed, remote sensing offers a number of independent means with which to retrieve various components of the hydrological cycle (e.g. rainfall, soil moisture, evaporation, terrestrial storage).

Page 1, line 27 Modified "observations" to "observation"

Page 2, line 3 regional- -to -global--scale

Page 2, line 4 Removed "key"

Page 2, line 5 Modified "is to characterize" to "is how to characterize"

Page 2, lines 6-8 Moved the following text to second paragraph in revised Introduction

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).

Page 2, lines 10-19 Modified in its entirety to the text below, and moved two paragraphs below (the following text constitutes the entirety of the fourth paragraph in the revised Introduction):

Observation-only studies are important, as they provide an unbiased perspective not just on hydrological closure, but also allow for a first-order examination of the underlying agreement between component variables. However, rather than just comparing the uncertainties between evaporation products and other hydrological components (which are poorly defined), there is still a need for alternative assessment techniques that explore the connection between the hydrological variables at both temporal and spatial scales. One approach to determine this is to evaluate the hydrological consistency between observed products (McCabe et al., 2008). The term hydrological consistency refers to the spatial and temporal match that should inherently exist between independent observations of hydrological states and fluxes, based upon physical considerations. It is a concept that encompasses the expectation of water cycle behavior and mass balance: that is, changes in one term should be reflected in related variables, both spatially and temporally. For instance, a rainfall event should result in an observable change in soil water storage and a consequent increase in evaporative flux, which in turn should reduce the available soil moisture. This relatively simple concept has been explored in the past, including in efforts to improve precipitation events by employing cloud detection methodologies (Milewski et al., 2009); using soil moisture changes to infer precipitation amounts (Brocca et al., 2014); examining the connection between soil moisture state and changes in atmospheric variables such as humidity and sensible heat flux (McCabe et al., 2008); as well as in assessments of land–atmosphere coupling between observations and reanalysis data (Ferguson and Wood, 2010).

Page 2, lines 19-24 Modified and moved to next paragraph (this new text constitutes the entirety of the third paragraph in the revised Introduction):

To~~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 (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.~~ this end, ~~A~~ a limited number of studies have sought ~~evaluated~~ to quantify large-scale ~~the~~ water budgets ~~closure of large basins across different regions of the world~~ using either satellite observations alone (Sheffield et al., 2009) or ~~via~~ through 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~~some of these studies (~~of~~ Sheffield et al., ~~(~~2009; Gao et al., 2010) ~~was to~~ evaluate ~~the~~ water budget closure ~~(~~by comparing the residual of the water budget (i.e. inferred runoff) with measured ~~streamflow values),~~runoff, ~~the remaining~~other studies ~~mostly~~ aim to provide merged or observation ~~/~~constrained estimates of the water cycle components~~, and any~~with estimates of uncertainty ~~is~~ given in terms of the variability among the products (e.g. Long et al., 2014). The results of these studies have generally illustrated large water budget closure errors, focusing on the temporal scale and invoking the use of a hydrological model to guide analysis or force closure, rather than being solely observation driven assessments.

Page 2, lines 24-31 Moved and modified (including additions to the text) the following text to fifth paragraph (this new text constitutes the entirety of the new fifth paragraph) in the revised Introduction.

In considering these earlier contributions, there remains a need to~~to~~ determine whether the ~~goal~~ basic idea of hydrological consistency can be realistically ~~achieved~~ extended to explore the agreement between ~~using~~ independent global-scale ~~currently available~~ satellite-based hydrological products. ~~In~~ To examine~~ing~~ this question, it makes sense to focus on ~~regions~~ catchments ~~with~~ that have relatively simple ~~water budgets~~hydrological interactions, as they represent natural laboratories within which the evaluation of large-scale products and the concept of hydrological consistency could be reasonably undertaken. For example, Wang et al. (2014) evaluated the level of agreement between three satellite-based hydrological cycle variables over arid regions in Australia, where surface and subsurface runoff were minimal. Given a sufficiently low runoff component, a lack of snow accumulation, and a relatively strong coupling of precipitation and evaporation components, arid and semi-arid environments represent potential candidates within which to undertake such process assessments. Recognizing the need to embrace a more holistic evaluation strategy, this study seeks to explore the hydrological consistency within a number of basins where hydrological processes are relatively simple, i.e. given the conditions described above. Our analysis constitutes a framework for assessing the utility of hydrological consistency to evaluate remotely sensed hydrological products~~, with a focus here being placed on an assessment of recently developed global evaporation products (McCabe et al., 2016; Miralles et al., 2016)~~. We undertake our analysis over four large river basins within arid and semi-arid environments distributed across the globe: the Colorado River basin in North America, the Niger River basin in Africa, the Aral Sea basin in Asia, and the Lake Eyre basin in Australia. ~~The rationale behind the selection of these study basins and the potential deviations from the ideal conditions of a relatively simple hydrological system are explained in Section 2. , 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.~~

Page 2, lines 31 – 33 Moved and modified (including additions to the text) the following text to sixth paragraph (this new text constitutes the entirety of the new sixth paragraph) in the revised Introduction.

In compiling datasets with which to evaluate and differentiate the three evaporation products, a number of product specific considerations needed to be accounted for~~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 estimates, which comprise~~ 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). As any continuous function on a sphere, the gravity field can be represented as an expansion in spherical harmonics, which form a complete set of basis functions in the sphere: similar to the way a Fourier series expansion uses sines and cosines as basis functions. Unlike precipitation and evaporation products (and most other hydrological remote sensing variables), it is problematic to directly compare spatial maps of GRACE water storage data with other spatially distributed hydrological variables (Tapley et al., 2004a), since GRACE data are usually filtered in the spectral domain. While scaling the GRACE data to account for differences due to filtering has been proposed as an alternative solution to this problem (Landerer and Swenson, 2012), it has recently been shown to affect results in some cases (Long et al., 2015), including arid regions. Given this restriction, we implement an alternative approach in which the precipitation and evaporation fields are transformed into spherical harmonics in order to remove the impact (and model dependence) of this scaling term. This approach allows for a more reasonable and equivalent intercomparison of hydrological variables, and represents a key novel aspect of this work. Further details describing this process are presented in Section 3. ~~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). 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).~~

Page 2, line 33 – Page 3, line 4 Moved to beginning of subsection 2.1 (GRACE water storage anomalies)

Page 3, lines 6 – 19 Modified (see following changes below) and moved in its entirety to second paragraph in revised Introduction

Page 3, line 6 Modified “serves as a key component in our analysis.” to:

“plays a key role in the water cycle as a linking mechanism between the surface and the atmosphere (Mueller et al., 2011)”

Page 3, line 7 Modified “or atmosphere that can be used” to “or atmosphere, which can be used”

Page 3, line 10 Modified “global scale” to “global-scale”

Page 3, lines 12 – 19 Modified text, including addition of moved text from Page 2, lines 6-8:

When ground-based flux observations are available, ~~these~~they can be used for calibration and evaluation:~~,~~ but large-scale assessment is inevitably constrained by the lack of distributed and representative in situ networks to comprehensively assess simulations (Jana et al., 2016) ~~,~~ as well as the inherent uncertainty ~~of~~associated with these observations. Some recent evaluation efforts have sought to estimate the uncertainty of satellite-based models of evaporation, as well as those from land surface model and reanalysis data, in terms of the variance amongst the products (Mueller et al., 2011; Jimenez et al., 2011; Long et al., 2014). These and related attempts have shown that no single evaporation product consistently outperforms any other, whether applied at local (Ershadi et al., 2014) or global scales (Miralles et al., 2016). Considering this issue of spatial mismatch and model variability, it seems inappropriate to assess these large-scale products via direct comparison to in situ data alone. Moreover, the quality of any satellite-based product should not be judged solely on its agreement with potentially unrepresentative point-scale approaches. Central 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 ground- and satellite-~~-~~based observations (McCabe et al. 2006). To compensate for this, it is important that a range of methods be used to evaluate the large-scale implementation of ~~these~~evaporation models.

Page 3, lines 6-19 End of changes to this paragraph, which has been moved in its entirety to second paragraph in revised Introduction.

Page 3, lines 21 – 32 Modified in its entirety to:

The o~~O~~verall objective of this study~~, the purpose of this paper~~ is to evaluate the hydrological consistency of three~~independent~~ satellite-based evaporation products with remotely sensed retrievals of precipitation and terrestrial water storage across a selection of basins that exhibit relatively well defined hydrological interactions. Throughout this analysis we aim to determine whether the hydrological consistency concept can expand the range of evaluation metrics used to assess large-scale hydrological data sets such as evaporation, and enable some differentiation amongstof relative product quality them to be made. 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 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 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.

Data and Methodology

Page 4, line 4 Modified "evaporation products" to "recently developed global-scale evaporation products"

Page 4, line 8 Inserted text moved from Page 2, line 33 – Page 3, line 4

Page 4, line 8 Modified "derived from " to "computed using"

Page 4, line 9 Inserted the following text:

The gravity field is usually described in terms of the geoid,: an equipotential surface that is defined to correspond to the mean sea level over the oceans (Swenson and Wahr, 2002). The geoid is usually approximated as a linear combination of spherical harmonics, given that these represent solutions to the Laplace equation that describes the relation between the gravitational potential and the geoid. The approximation is of the form:

Page 4, line 10 Inserted equation moved from Page 7, line 18 (i.e. equation 1)

Page 4, line 11 Inserted and modified text from Page 7, lines 19-22:

where $\tilde{P}_{lm}(\cos\theta)$ are the normalized associated Legendre functions, θ corresponds to the colatitude (the complementary angle to the latitude), ϕ to longitude, C_{lm} and S_{lm} are the spherical harmonic coefficients of degree l and order m, and l_{max} is the truncation degree. The total number of coefficients is given by $((l_{max}+1)^2 + l_{max}+1)/2$, while the resolution (the scale of the smallest feature of the gravity field that can be resolved using l_{max} coefficients) is approximately $\pi a/l_{max}$ (where a is the Earth's radius). The data product used in this study (processed by UTCSR) contains coefficients up to $l_{max} = 60$, i.e. a total number of 1891 coefficients, with an approximate resolution of 333 km. The full description of the process to transform the gravity field anomalies into water storage anomalies is described in detail in Wahr (1998) and Swenson and Wahr (2002).

Page 4, lines 9-20 Modified text and moved to next paragraph (following text constitutes the new paragraph in its entirety):

GRACE data contains two types of errors (correlated and random) that need to be filtered before translating the data into water storage anomalies. Correlated errors are known to contaminate the signal in the form of north-south oriented stripes. A "de-striping" filter was applied to the coefficients (Swenson and Wahr, 2006; Duan et al, 2009) in order to remove this source of error. An isotropic filter (Gaussian filter with radius of 300 km) was then used to remove random errors (Swenson and Wahr, 2002). Furthermore, it is a usual practice to replace the The degree 2 coefficients with a more reliable estimate from a low-degree model of the gravity field calculated using 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 filterfilters haves 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 surface model (LSM). This method is has been used to obtain scaling factors for GRACE data in order to restore the signal (it has been observed that the filters typically reduce the signal) before using the GRACE data with other hydrological variables (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 element of uncertainty in our study, which is focused on arid regions, we instead transformed the other water cycle components (i.e. evaporation and precipitation) into spherical harmonics, using an approximation similar to equation 1. The effect of the filters is therefore incorporated directly into the study by applying the filters to the variousother hydrological components in spherical harmonics. (see Section 3.1).

Page 4, line 26 Modified "described briefly" to "briefly described"

Page 5, line 3 Modified "Leaf Area Index" to "leaf are index"

Page 5, line 4 Removed "and Normalized Difference Vegetation Index (NDVI),"

Page 5, line 5 Added ", a three--source scheme used for terrestrial land flux estimation." after "(MOD16)"

Page 5, line 7-8 Modified "Other improvements include:" to "Other adjustments incorporated into MOD16 include"

Page 5, lines 9-10 Modified "The product includes transpiration and evaporation from soil and wet canopy," to "The MOD16 product includes comprises transpiration, and evaporation from the soil and wet canopy,"

Page 5, lines 10-11 Removed "Potential evaporation (calculated using a Priestley-Taylor based formulation) is also included to monitor environmental water stresses and droughts (Mu et al., 2011)"

Page 5, line 11 Added new text:

Each component is weighted-based on the fractional vegetation cover, relative surface wetness and available energy. Inputs to the model include net radiation (Rn), air temperature and humidity, as well as LAI and vegetation phenology. Importantly, it does not require wind speed or soil moisture data, making it a relatively parsimonious model in terms of input requirements. In this study, we used the actual evaporation (AET) product from MOD16 (Mu et al., 2011) with 8-day temporal resolution and 1 km resolution in the sinusoidal projection. The product was reprojected onto a 0.05° regular grid using the MODIS Reprojection Tool (MRT) before transformation into spherical harmonics, as described in Section 3.1. Further details on the modeling basis behind the MOD16 product can be found in Mu et al. (2013), Ershadi et al. (2014) and Michel et al. (2016).

Page 5, line 14 Modified “resulting in the Penman-Monteith-Leuning (PML) model” to “resulting in the two-source Penman–-Monteith–-Leuning (PML) model”

Page 5, line 15 Modified “canopy conductance.” to “the ~~canopy conductance~~surface resistance, which was previously calculated as LAI multiplied by a constant c_L (Cleugh et al., 2007).”

Page 5, lines 15-19 Removed completely and replaced by the following new addition:

The new parameterization of the surface resistance in the PML model was optimized using data from 15 globally distributed flux station sites, with two key parameters identified: the maximum stomatal conductance (gsx) and the ratio of actual to potential evaporation at the soil surface. Zhang et al. (2010) developed a method to further optimize the spatial variability of these two parameters (i.e. at each grid pixel) using gridded meteorological data and a simple Budyko-curve hydrometeorological model developed by Fu (1981) that includes precipitation and available energy as inputs. Mean annual evaporation for each grid pixel is calculated using the Fu model and gridded meteorological data. The value of gsx is optimized using a non-linear least square regression-based on the difference between the PML and the Fu model. Interestingly, the Fu model is calibrated by comparing the output evaporation with the residual of precipitation and runoff i.e. by assuming negligible annual water storage changes and groundwater inflow and outflow. Zhang et al. (2012) used this approach to develop a global gridded terrestrial evaporation product (hereafter referred to as CSIRO-PML; Zhang, 2014, personal communication) with a 0.25° resolution (in this study, we used the actual evaporation product). They used gridded meteorological data from diverse sources, including vapor pressure and temperature from the Climate Research Unit (New et al., 2000), LAI and land cover type from Boston University (Ganguly et al., 2008), precipitation from the Global Precipitation Climatology Centre (GPCC, version 4; Rudolf and Schneider 2004), and radiation data from the Global Energy and Water Cycle Exchanges (GEWEX) Surface Radiation Budget (Gupta et al., 2006).

Page 5, line 21 Modified “satellite data based” to “satellite-based”

Page 5, line 23-26 Removed all text after “(Gash, 1979)”, replaced by:

(Gash, 1979) as a first step. GLEAM then employs the Priestley–Taylor equation to calculate the potential evaporation of bare soil and vegetation components (both short and tall canopy), with values constrained to actual evaporation via application of a stress factor. The stress factor is calculated using vegetation optical depth from a combination of different satellite passive microwave observations using the Land Parameter Retrieval Model (Liu et al. 2013). GLEAM also has the capacity to explicitly calculate sublimation of snow covered surfaces (Takala et al., 2011) as well as open water evaporation. Satellite observations of surface soil moisture can be assimilated using a Kalman filter assimilation approach to estimate the moisture profile over several soil layers. Here we employ version 2A of GLEAM (Miralles, 2014, personal communication), which uses a combination of satellite, ground and reanalysis input data. Precipitation is obtained from the Climate Prediction Center Unified data set, consisting of data from over 30,000 stations (CPC-Unified, Joyce et al., 2004). The radiation product used in this version of GLEAM is the European Center for Medium-Range Weather Forecasts (ECMWF) ERA-Interim meteorological reanalysis product (Dee et al., 2011). In this version of GLEAM, surface soil moisture data from the Water Cycle observation Multi-mission Strategy Climate

Change Initiative (WACMOS-CCI) merged product (from a combination of several passive and active microwave products) is assimilated (Liu et al., 2012), while air temperature is derived from both the International Satellite Cloud Climatology Project (ISCCP) and the Atmospheric Infrared Sounder (AIRS) (Rossow and Dueñas, 2011). Further details of the model can be found in Miralles et al. (2010), Miralles et al. (2011a) and Miralles et al. (2011b).

Note: Due to the addition of a new citation "Miralles et al. (2011b)" in the text above, all previous citations to "Miralles et al. (2011)" have been replaced to "Miralles et al. (2011a)"

Page 6, line 5 Inserted "and has been previously used in soil moisture- and evaporation-based analyses" before "(Crow, 2007; Miralles et al., 2011a)"

Page 6, lines 9-10 Modified text:

While runoff data was not used explicitly in the consistency analysis presented in this manuscript, ~~Although not considered in the evaluation of hydrological consistency,~~simulated runoff data was~~is~~ compared to precipitation and evaporation observations~~data~~ in order to evaluate ~~whether~~ the assumption of a relatively simple ~~water budget~~hydrological system ~~is valid~~ in the study ~~regions~~basins.

Page 5, line 14 Inserted following text before "The version of the product…"

Although these values were not constrained with ground estimates and thus may contain biases, as noted, runoff values were only used to provide an assessment of runoff against the observed precipitation and evaporation data.

Page 6, line 16 Modified "2.5 Study basins" to "2.5 Selection of study basins"

Page 6, lines 17-18 Modified text:

The study basins~~regions~~ were ~~chosen~~ targeted ~~targeted~~ primarily~~mainly~~ on their climate classification, with river basins in~~that covered~~ regions with a predominantly arid or semi-arid climate preferentially selected.

Page 6, line 18 Inserted text before "We employed…":

This criterion was established in order to seek a relatively simple hydrological system (i.e. constrain the range of possible hydrological interactions), thereby maximizing the conditions under which hydrological consistency between evaporation and precipitation and water storage changes might be achieved.

Page 6, line 18 Modified "We employed a Köppen classification map" to "A Köppen classification map,"

Page 6, line 19 Inserted text ", was used to identify arid and semi-arid regions." after "(Kottek et al., 2006)".

Page 6, lines 19-20 Moved "The climate criteria was to select basins with more than 50% areal extent containing any of the arid Köppen climates (BWk, BWh, BSk or BSh)." after "…(USGS)" (Page 6, line 23)

Page 6, lines 21-22 Modified "(GRDC), derived" to "(GRDC) and derived"

Page 6, line 23 Inserted text before "Secondary"

A threshold of 50% areal extent containing any of the arid Köppen climates (BWk, BWh, BSk or BSh) was used to select potential basins.

Page 6, lines 23-24 Modified text:

Secondary ~~considerations~~ criteria for ~~selecting~~ basin selections from the GRDC data set ~~were based~~focused on size, geographical ~~location~~ distribution and amplitude and trends in the water storage variations.

Page 6, lines 24-25 Modified text:

In terms of the size of the basin, a smaller size would more likely satisfy the assumption of a relatively simple hydrological system. However, due to the coarse resolution of GRACE data (see section 2.1), this requirement had to be compromised. Given these considerations, ~~We selected~~ four basins were selected as focus regions of study: the Colorado R ~~R~~iver basin (CRB) in North America, the Niger River basin (NRB) in Africa, the Aral Sea basin (ASB) in Asia and the Lake Eyre basin (LEB) in Australia (Figure 1).

Page 6, line 29 Modified “compare” to “establish”

Page 6, line 29 Modified “, i.e. a simple water budget,” to “(i.e. a simple water budget)”

Page 6, line 30 Modified “pre-dominantly arid” to “predominantly arid,”

Page 7, line 2 Modified “Similarly” to “Likewise”

Page 7, line 4 Modified “as important as evaporation” to “that is ~~similar in~~ close in magnitude to~~as important as~~ evaporation”

Page 7, line 5 Added “undertaken here.” after “analysis”

Page 7, line 6 Inserted text, including modified text moved from page 10, lines 6-7, 22-23, and page 11, lines 9-10, 25-26.

Even though these four basins were preselected based upon their location within dryland systems (Wang et al., 2012), they reflect a range of trends in water storage and precipitation. For example, the Colorado River basin experienced intervals of wet and dry periods, while the Niger River basin showed a small but steady increase in water storage with a clear, seasonal variability in both water storage and precipitation. Meanwhile, the Aral Sea basin experienced a significant loss of water during the study period ($-8.2\ mm.yr^{-1}$), in line with the historical depletion of this inland sea in response to increased agricultural productivity (Zmijewski and Becker, 2014). ~~Finally, t~~The Lake Eyre basin showed a marked increase in precipitation during the end of the study period (2009-2011), with a corresponding increase in water storage during the following years, reflecting the larger scale hydrometeorological conditions affecting that region (Boening et al., 2012).

Methodology

Page 7, line 8 Inserted “and to ensure that a fair comparison between GRACE data and satellite products could be undertaken” after “(i.e. at sub-basin scale)”

Page 7, line 8 Modified “Futhermore, the” to “The”

Page 7, line 8 Inserted “(see Section 2.1)” after “de-striping filter”

Page 7, line 8 Modified “accounted for” to “incorporated into the analysis”

Page 7, lines 13-22 Moved and modified to Page 4, line 10 (Section 2.1 GRACE water storage anomalies)

Page 7, line 24 Inserted “up to an approximation of degree l_{max}” after “C_{lm} and S_{lm}”

Page 7, line 25 Modified “In this study,” to “Here”

Page 7, lines 29-30 Removed line break

Page 8, line 1 Modified “for each E data set” to “for each E product”

Page 8, line 2 Inserted “(see section 2.1)” after “GRACE TWSA data”.

Page 8, line 4 Modified “The computed spherical harmonic coefficients so far are global” to “In the analysis so far, the computed spherical harmonic coefficients are global”

Page 8, line 5 Modified “needed to be” to “needs”, and “regions” to “basins”

Page 8, line 8 Modified “at a basin-scale” to “at the basin scale”

Page 8, line 27 Modified “region” to “basin”

Results

Page 9, line 3-5 Removed “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?”

Page 9, line 5 Modified “A cursory examination” to “An examination”

Page 9, line 21 Modified “hydrological closure approach” to “hydrological consistency approach”

Page 9, line 22 Added quotes to “simple”

Page 9, line 22 Added text before “While it is …”:

> Indeed, this has been demonstrated in other studies using either satellite data alone, or a combination of satellite and ground data.

Page 9, line 22 Modified “this current work” to “the current work”

Page 9, line 23 Modified “these different evaporation models” to “these different evaporation models based on hydrological closure”

Page 9, line 24 Added “(precipitation and gravity-based water storage changes) that the evaporation is” before “being compared against”

Page 9, lines 27-28 Removed “A key objective of this analysis is to assess the hydrological consistency between discrete components of the water cycle (see Figure 2).”

Page 9, line 28 Modified "To do this," to "In this section," and "examined" to "examine"

Page 9, line 29 Modified "Figures 4-7" to "Figure 4 to Figure 7"

Page 9, line 30 Removed "forming the focus of this study"

Page 10, line 1 Inserted "e.g. . For example, do the degree correlations behave differently during wet or dry periods, or when storage changes can be attributed to eitherare causeddriven by natural or anthropogenic causes?"

Page 10, lines 1-2 Removed "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?"

Page 10, line 5 Modified "4.2.1 Colorado river basin (CRB)" to "4.2.1 Colorado River basin"

Note: throughout the text, we replaced all basin abbreviations (e.g. CRB as in Colorado River basin) to the full name of the basins. Only the first mention of the abbreviations were keft (i.e. in Section 2.5) as well as in the Figures.

Page 10, lines 6-7 Modified text

> The start of the study period (2003) coincided with the end of an intense multi-year drought in the Colorado River basin (Scanlon et al., 2015).

Page 10, line 10 Corrected "February 2004" to "February 2005"

Page 10, line 16 Modified "runoff in the basin: a large" to "runoff in the basin, since a large"

Page 10, line 19 Modified "Figure 4, however" to "Figure 4. However,"

Page 10, line 22 Modified "region" to "basin", "and a clear" to "with clear". Added "(see Figure 5)" after "seasonality".

Page 10, lines 23-25 Modified text:

> Over the study period, pPrecipitation peaked peaks between July and September, while TWSA peaked peaks between September and November. Ahmed et al. (2014) attributed the observed increase in TWSA to an increase in precipitation in the region caused by warmer Atlantic Ocean temperatures.

Page 10, line 28 Modified "GRACE observing period, the region" to "GRACE observing period (2003-2004), the basin"

Page 11, lines 5-7 Modified text:

> Interestingly, there also seems to be less inter-degree variability compared to the other regionsstudy basins. This may be related to the simpler water budget in this region basin compared to that of the Colorado rRiver and Aral Sea basinRB ands ASB, but requires further investigation.

Page 11, line 9-11 Modified text:

> The Aral Sea basin is an is endorheic basin that has experiencedcontinued a historical trend of water loss a significant loss of water during the study period, most likely caused by agricultural

activities ((8.2 mm.yr^{-1}), in line with the historical depletion of this inland sea in response to increased agricultural productivity (Zmijewski and Becker, 2014).

Page 11, line 15 Added "(see Figure 6)" after "respectively"

Page 11, line 18 Modified "region" to "basin"

Page 11, lines 21-23 Modified text:

> These complications are reflected in the higher inter-degree variability (compared to the other regionsbasins), as well as in the slight differences in degree correlation between the evaporation products.. Differences in degree correlation due to the use of the three evaporation products were minimal i.e. no single evaporation product resulted in a significantly higher (or lower) hydrological consistency with precipitation and water storage anomalies.

Page 11, line 25 Modified "rainy seasons in" to "rainy seasons of"

Page 11, line 28 Added "(see Figure 7)" after "period"

Page 12, line 1 Modified "In general,“ to "Overall,"

Page 12, line 3 Added text after "spatial and temporal patterns. "

> More importantly, none of the evaporation products showed a significant (and persistent) advantage in terms of hydrological consistency over the others.

Page 12, line 4 Modified text

> (TRMM--based P, MOD16--based E and GRACE TWSA) over the LEBthis basin,

Page 12, lines 7-15 Modified text

> Before GRACE can detect a significant water storage increase, the water mass from precipitation needs to build reachup 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, during which as water accumulates in the soil and travels from different source 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 from amongst thethose basins studied here is in the Niger River basinRB (Figure 5), where a lag of two months is evident throughout the study period. In other regions, however, such as the Colorado River basin RB and Lake Eyre basinB, 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 geomorphological characteristics (Ahmed et al., 2011; Wang et al., 2014). Because of their large extent and geographical features, the Colorado river and Aral Sea basins include regions where snow storage plays an important role as delayed sources of runoff. Other potential sources of delayed flow, not limited to these two basins, include groundwater flow and surface water flow. The combination of these slowly varying components of streamflow is defined as baseflow (Beck et al., 2013).

Page 12, lines 17-22 Modified text

To examine this temporal component, at least in a simplified manner,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 Niger River RbasinB 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 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). Beck et al. (2013) developed global estimates of the Base Flow Index (BFI):, a measure of the ratio of the long-term baseflow to the total runoff, using a large global data set of runoff and a regionalization procedure to transfer these and other characteristics of runoff from gauged to ungauged basins. BecauseSince we did not model any of the physical processes contributing to baseflow, the BFI couldwas examined to assist in help explaining part of the delay in observed water storage changes relative to the P-E term (, although not dynamically since (the index is a long-term average in time)). The spatial average of the BFI in the four study basins is, not surprisingly, within the same range: between 0.4 and 0.6. This is not unexpected, as various climate characteristics were used as predictors of BFI. FurthermoreIndeed, the fact that they are similar is in agreement with our finding of similar GRACE lag times among the study basins. Further investigation is required to determine the nature of the elements affecting the lag in water storage, not limited to those found in baseflow.

Discussion

Page 13, lines 5-22 Modified text and added new text to paragraph:

To date, tThe development of methods and sensors to retrieve the various components of the water cycle has for the most partlargely been undertaken independently of other (interrelated) variables processes (see McCabe et al., 2008 and Brocca et al., 2014 for some examples of complementary retrieval). Large-scale retrievals of hydrological variables such as evaporation, soil moisture and rainfall products do not come with well-defined accuracy metrics, let alone uncertainty bounds. TheThis lack of any well-defined error structure associated with individual products complicates the task of product assessment. As such, the question of how to evaluate large-scale thesdatasets remains an outstanding one. This is especially important in the context of global-scale products. While a number of global evaporation (and precipitation) evaluation papers have been published, none seek to identify consistency with related hydrological variables, and focus instead on comparisons against traditional point-scale or tower-based techniques (McCabe et al., 2016; Michel et al., 2016). Given the spatial mismatch between ground observations (and the lack of continuous large-scale coverage of in situ data in remote regions), it is inappropriate to evaluate these large-scale products in such a manner. Determining whether individual products are at the least consistent with each other (i.e., they reflect hydrological expectation) is a needed first step in product assessment. The motivation behind this study was to take a step back and determine whether a first order hydrological assessment could be achieved. Rather than comparing the uncertainties between the evaporation products and the other hydrological components (which are poorly defined), we attempt to distinguish between the different evaporation products relative to their consistency with precipitation and terrestrial water storage. That is, are observed changes or patterns in the evaporation datasets reflected in these other hydrological variables? We explore this approach precisely because of the challenges in quantifying uncertainty based upon traditional in situ methods. As is discussed below, the challenge on how to do this remains, raising some important questions on both product quality and also the techniques we use to evaluate global products.

Page 13, lines 10-11 Modified and moved to next paragraph

It is, 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 inter-product agreement between hydrological components should existmight be found.

Page 13, lines 17-18 Modified and moved to next paragraph

Given the relationship between size and accuracy for GRACE data, a geographically distributioned selection of basins that could satisfyfit into this simplified water budget assumption is somewhat limitedwas difficult.

Page 13, line 23 Added new sub-heading "5.1 Challenges to implementing hydrological consistency"

Page 13, line 24 New paragraph based on modified text from Page 13, lines 10-11 and 17-18

It is, 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 inter-product agreement between hydrological components should existmight be found. For this reasonTo explore this idea, ourthe study assumed focused on basins where aa simpler water budget, consisting of water storage anomalies as a function of precipitation and evaporative fluxes, might be expected to predominate. The aim was to limit potentially complicating variables such as snow, vegetation changes, large precipitation and streamflow contributions and other hydrological processes. The assumption was that arid- and semi-arid regions would best fit this profile. 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 distributioned selection of basins that could satisfyfit into this simplified water budget assumption is somewhat limitedwas difficult. Restrictions related to basin size affect the study in two conflicting ways. On the one hand, a large basin will inevitably present complications related to heterogeneity (including in climate zones, as was the case for the Colorado River basin and Aral Ssea basin) and also be more likely to contain areas affected by anthropogenic activities, such as irrigation, land cover changes, building of dams and reservoirs, etc. On the other hand, a small catchment size would be more difficult to evaluate with this consistency approach, given the coarse resolution of (most) of the global products used here, but especially the GRACE data. The spatial resolution of GRACE data is further limited by the use of filters to remove errors. Considering these restrictions, a compromise in the selection of study basins was required to allow for at least a narrow range of length scales (500-800 km) to be evaluated.

Page 13, lines 18-22 Modified and moved as third paragraph in the revised Discussion section:

In the end, our study consisted of four major globally distributed river basins, including two endorheic systems. Although having mostly an arid climate in terms of Köppen classification, both the Colorado River and Aral Sea basins include regions with the presence of snow and snowmelt-dominated runoff. While snow storage itself is not a problem, since GRACE detects changes in storage irrespective of their nature (snow, groundwater, soil moisture, etc), snowmelt may contribute to delayed changes in storage that can affect gravity results. As such, the inclusion of these basins was considered important in order to test the hydrological consistency concept in regions that deviated from the ideal assumption. The influence that snowmelt and other potential sources of lag in the system have is poorly defined and forms part of the reason to explore the inclusion of a lag response in the GRACE data (see Section 4.3).

Page 13, lines 24 to Page 14, line 14 Removed

Page 13, Addition of a new paragraph (fourth) based on text from Page 15, lines 12-23:

Apart from the issues of spatial scale, the use of satellite-based hydrological data presents additional challenges and sources of uncertainty to a consistency-based assessment. For instance, because GRACE data is smoothed to remove errors in short-scale terms (i.e., truncation of the spherical harmonic coefficients), the gravity signal contains contamination from outside of the studied basins (leakage) and is a potential source of uncertainty in areas neighbouring high amplitude signals (particularly if they are out of phase~~single~~ ~~signal~~with the study basin) and the ocean. Although the LSM-based scaling factor, which is static in time, has been used to correct for bias (e.g. signal reduction) and leakage contamination, dynamic changes in water storage trends outside the basin might still contaminate the signal (Long et al., 2015)~~-~~. In addition, the temporal lag in terrestrial storage response as documented in previous studies (Rieser et al., 2010; Ahmed et al., 2011; Wang et al., 2014; Ayman and Jin, 2016) and observed in our analysis, ~~was~~represents an important source of potential error (see Sections 4.3 and 5.2). Product errors are also evident in the precipitation and evaporation data sets. Global rainfall retrievals have well recognised limitations, including the detection of both high and low intensity events (Hou et al., 2014), the discrimination of cloud clear and cloud precipitation scenes, as well as the sensitivity to parameters in the forward model of radiative transfer over different sensors (Stephens and Kummerow, 2007). In terms of evaporation, uncertainties related to algorithm choice, input data variability and process parameterizations all complicate the accurate estimation of terrestrial fluxes (Ershadi et al. 2015). Determining whether or not and understanding how much these product sources of uncertainty affect hydrological consistency studies remains an area requiring further investigation.

Page 14, lines 14- 26 Modified text; moved to new fifth paragraph under a new sub-heading "5.2 Temporal lag in terrestrial storage response":

In exploring the relationship between GRACE water storage changes and precipitation and evaporation data, it was evident that water storage anomalies peaked at a significantly later time than the corresponding P-E values. One possibility for this apparent lag in ~~that has been explored in studies using~~ GRACE data ~~(Rieser et al., 2010; Ahmed et al., 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 (a rough estimate of GRACE accuracy averaged over the entire Earth is 20 mm.month^{-1}; Wahr and Velicogna, 2006),~~,~~ ~~and therefore~~ the corresponding mass is not detected until a sufficient amount has accumulated within the catchment via natural drainage processes (Rieser et al., 2010; Ahmed et al., 2011)~~.~~. 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 whole study period. Doing t~~T~~his ~~seemed to~~ improved 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 ~~r~~R~~R~~iver 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 R~~r~~~~R~~iver basin and Aral Sea basin, where changing trends in the seasonal patterns of precipitation ~~make~~ made 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 effect on the degree correlation. Further understanding the implications and physical rationale behind the attribution of this lag is required.

Page 14, line 28 to Page 15, line 10 Modified text; moved to new sixth paragraph under a new sub-heading "5.3 Discriminating between satellite evaporation products"

~~A secondary objective~~One aspect of this work was to ~~determine~~ explore whether differences in available evaporation products could affect the results of the consistency analysis i.e. could we identify ~~was there~~ better agreement between water storage anomalies and P-E ~~with~~ in any~~one~~ particular evaporation product? While t~~T~~he analysed products ~~used in the study~~ covered a wide range of resolutions (0.05°–0.25°), ~~although~~ the effective resolution in the analysis ~~is~~ was ultimately determined by the truncation degree (l_{max}) of the spherical harmonic transformation. Even after accounting for this, absolute differences were evident from a qualitative basin-scale analysis (Figure 2). ~~Overall, r~~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 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. Whether these differences in model parameterization are the sole cause of the apparent underestimation by MOD16 remains to be investigated.

Page 15, lines 12-23 Moved and extensively modified to serve as the fourth paragraph in the revised Discussion Section (see text above)

Page 15, line 25 to Page 16, line 7 Extensively modified and separated into two new paragraphs:

~~Because~~As the focus of the study was to discriminate between evaporation products, the question on whether the choice of precipitation product affects the hydrological consistency analysis ~~remains~~was somewhat beyond the scope of this work. However, a –first order analysis was undertaken by replacing the GPCP precipitation ~~field~~data with another data set and reproducing the analysis. To do this, we processed the Precipitation Estimation from Remote Sensing Information using Artificial Neural Network (PERSIANN) product, which uses an a~~A~~rtificial neural network to approximate spatiotemporal non-linear relationships between physical variables and remotely sensed signals (Hsu et al., 1997). PERSIANN uses data from the long wave infrared imager onboard the Geostationary Operational Environmental Satellite (GOES) as well as from the Tropical Measurement Mission (TRMM) microwave imager (TMI). As shown in Figure S1, ~~T~~the results of this new analysis did not ~~showcase~~reveal any significant difference ~~than~~when compared to those based on ~~with~~ the GPCP analysis. Figure S1 shows the average degree correlation statistics per study region and evaporation product, with and without the inclusion of a lag in GRACE data.

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

Evaluating global evaporation products remains an outstanding challenge. The purpose of implementing a hydrological consistency approach was to explore the evaluation of evaporative fluxes by comparing the spatial patterns between precipitation and changes in water storage. If such an approach could be shown to perform well in a relatively simple hydrological system, the potential for broader-scale application in regions with more complex behaviour would be the next logical step. However, the study showed that even in these relatively simple basins, it was not possible to demonstrate a consistent hydrological agreement between ~~current~~ independent observations. Improvements in satellite-based evaporation products ~~will~~are likely to be delivered 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~~anticipated to provide more accurate estimates of the water storage anomalies. Although products will inevitably improve (or continue to improve in an individual manner), it will ~~be~~remain important to evaluate them in a holistic manner. ~~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 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.~~

Conclusion

Page 16, lines 9-10 Modified “one of the motivating elements of this study” to “a key ~~one of the~~ motivation behind~~ng elements of~~ this study”

Page 16, line 10 Removed “(or estimates)”

Page 16, line 11 Modified “to achieve” to “to reflect”, and “closure” to “consistency”

Page 16, line 11 Modified “To do this” to “To do so”

Page 16, line 12 Modified “hydrological consistency” to “this”

Page 16, line 12 Modified “that is, in arid regions” to “in arid and semi-arid regions”

Page 16, line 12 Modified “water budget consisting (ideally) of only” to “water budget, consisting primarily of”

Page 16, line 13 Modified “We found” to “Unfortunately, we found”

Page 16, line 14 Removed “throughout the study period”

Page 16, line 16 Modified “in individual products” to “within individual products”

Page 16, line 16 Removed “Furthermore,”

Page 16, line 17 Modified “significant differences” to “significant and known differences”

Page 16, line 18 Modified “did not play” to “did not seem to play”

Page 16, lines 18-20 Moved and modified text to next paragraph:

> While ~~Although~~ not providing a comprehensive tool for product evaluation, the approach did help to reveal some interesting spatial and temporal patterns between the studied hydrological variables.

Page 16, line 25 Modified "then" to "with"

Page 16, line 27 Added "quite" before "reasonable"

Page 16, line 28 Added "to account for delayed sources of water storage changes" after "considered"

Page 17, line 1 Modified "in other regions" to "in some of the studied basins"

Page 17, line 7 Modified "as a potential tool in" to "as an element of a holistic approach to ~~potential tool in~~"

Page 17, line 26 Added new references:

Beck, H. E., Dijk, A. I. J. M. van, Miralles, D. G., Jeu, R. A. M. de, Bruijnzeel, L. A. (Sampurno), McVicar, T. R. and Schellekens, J.: Global patterns in base flow index and recession based on streamflow observations from 3394 catchments, Water Resources Research, 49(12), 7843-7863, doi:10.1002/2013WR013918, 2013.

Boening, C., Willis, J. K., Landerer, F. W., Nerem, R. S. and Fasullo, J.: The 2011 la niña: So strong, the oceans fell, Geophysical Research Letters, 39(19), doi:10.1029/2012GL053055, 2012.

Brocca, L., Ciabatta, L., Massari, C., Moramarco, T., Hahn, S., Hasenauer, S., Kidd, R., Dorigo, W., Wagner, W. and Levizzani, V.: Soil as a natural rain gauge: Estimating global rainfall from satellite soil moisture data, Journal of Geophysical Research: Atmospheres, 119(9), 5128-5141, doi:10.1002/2014JD021489, 2014.

Page 18, line 5 Added new references:

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., Berg, L. van de, Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., Rosnay, P. de, Tavolato, C., Thépaut, J.-N. and Vitart, F.: The ERA-interim reanalysis: Configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137(656), 553-597, doi:10.1002/qj.828, 2011.

Page 18, line 20 Added new references:

Ferguson, C. R. and Wood, E. F.: Observed land-Atmosphere coupling from satellite remote sensing and reanalysis, Journal of Hydrometeorology, 12(6), 1221-1254, doi:10.1175/2011JHM1380.1, 2011.

Fu, B. P.: On the calculation of the evaporation from land surface (in Chinese), Sci. Atmos. Sin., 5, 23–31, 1981.

Gao, H., Tang, Q., Ferguson, C. R., Wood, E. F. and Lettenmaier, D. P.: Estimating the water budget of major US river basins via remote sensing, International Journal of Remote Sensing, 31(14), 3955-3978, doi:10.1080/01431161.2010.483488, 2010.

Page 18, line 31 Added new references:

Hsu, K.-l., Gao, X., Sorooshian, S. and Gupta, H. V.: Precipitation estimation from remotely sensed information using artificial neural networks, Journal of Applied Meteorology, 36(9), 1176-1190, 1997.

Jana, R. B., Ershadi, A. and McCabe, M. F.: Examining the relationship between intermediate scale soil moisture and terrestrial evaporation within a semi-arid grassland, Hydrology and Earth System Sciences Discussions, 2016, 1-27, doi:10.5194/hess-2016-186, 2016.

Page 19, line 13 Added new references:

Liu, Y. Y., Dijk, A. I. J. M. van, McCabe, M. F., Evans, J. P. and Jeu, R. A. M. de: Global vegetation biomass change (1988-2008) and attribution to environmental and human drivers, Global Ecology and Biogeography, 22(6), 692-705, doi:10.1111/geb.12024, 2013.

Page 19, line 32 Added new references:

Miralles, D. G., Gash, J. H., Holmes, T. R. H., Jeu, R. A. M. de and Dolman, A. J.: Global canopy interception from satellite observations, Journal of Geophysical Research: Atmospheres, 115(D16), doi:10.1029/2009JD013530, 2010.

Page 19, line 34 Modified “2011” to “2011a”

Page 20, line 1 Added new references:

Miralles, D. G., De Jeu, R. A. M., Gash, J. H., Holmes, T. R. H. and Dolman, A. J.: Magnitude and variability of land evaporation and its components at the global scale, Hydrology and Earth System Sciences, 15(3), 967-981, doi:10.5194/hess-15-967-2011, 2011b.

Page 20, line 29 Added new references:

Rossow, W. B. and Dueñas, E. N.: The international satellite cloud climatology project (ISCCP) web site: An online resource for research, Bulletin of the American Meteorological Society, 85(2), 167-172, doi:10.1175/BAMS-85-2-167, 2004.

Page 21, line 22 Added new references:

Takala, M., K. Luojus, J. Pulliainen, C. Derksen, J. Lemmetyinen, J. Kärnä, J. Koskinen, and B. Bojkov.: Estimating Northern Hemisphere Snow Water Equivalent for Climate Research through Assimilation of Space-Borne Radiometer Data and Ground-Based Measurements, Remote Sensing of Environment. 115. 3517-3529. http://dx.doi.org/10.1016/j.rse.2011.08.014, 2011.

Page 22, line 1 Added new references:

Wang, L., D’Odorico, P., Evans, J. P., Eldridge, D. J., McCabe, M. F., Caylor, K. K. and King, E. G.: Dryland ecohydrology and climate change: Critical issues and technical advances, Hydrology and Earth System Sciences, 16(8), 2585-2603, doi:10.5194/hess-16-2585-2012, 2012.

Page 27, Table 1 Updated reference “Miralles et al. (2011)” to “Miralles et al. (2011a,b)”

Page 30, line 6 Added new supplementary Figure and caption:

Figure S1. 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. This figure represents a summary of the analysis using the PERSIANN product as precipitation. The results are very similar to those in Figure 8.

Evaluating the hydrological consistency of evaporation products using satellite~~-~~-~~-~~based gravity and rainfall data~~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~~space~~-~~-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~~develop regional~~-~~ -to -global~~-~~scale estimates of evaporation, ~~products have been developed,~~ offering insights into ~~the~~ this key component of the hydrological cycle.~~multi-scale behaviour and variability of hydrological states and fluxes.~~ However, the evaluation of ~~such~~ large-scale evaporation products is not a straightforward task. While a number of studies have intercompared a range of these products by examining the variance amongst them or by comparison of pixel-scale retrievals against ground-based observations, there is a need to explore more appropriate techniques to comprehensively evaluate ~~them~~flux estimates. ~~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.~~One possible approach is to establish the level of product agreement between related hydrological components: for instance, how well do evaporation patterns and response match with precipitation or water storage changes. ~~In order t~~To assess the suitability of this ~~type of~~"consistency"~~-~~-based approach for evaluating ~~hydrological observations~~evaporation products, we identified four globally distributed basins in arid- and semi-arid environments, ~~including~~~~comprisin~~including the Colorado River basin, the Niger River basin, the Aral Sea basin and the Lake Eyre basin. In an effort to identify product quality, three satellite-based global evaporation products including CSIRO-PML, MOD16 and GLEAM were evaluated against rainfall data from GPCP ~~along~~~~together~~along with GRACE water storage anomalies~~ in the form of spherical harmonic (SH) coefficients~~. To ensure a fair comparison, we evaluated consistency using a degree correlation approach after transforming both evaporation and precipitation data into spherical harmonics. Overall we found~~, 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 dryland environments. Indeed, the degree correlation showed oscillating values between periods of low and high water storage changes, with a phase difference of about 2-3 months.. i.e. no significant or persistent advantage was identified across any of the three evaporation products in terms of a sustained hydrological consistency with precipitation and water storage anomalies data. HoweverInterestingly, after imposing a simple lag in GRACE data to account for delayed surface runoff or baseflow components, there was an improved match in terms of degree correlation in the Niger River basin. Significant improvements (from ~0 to about 0.6) to the degree correlations (from ~0 to about 0.6) were also found in the Colorado River basins for both the CSIRO-PML and GLEAM products (MOD16 showed only half of that improvement). for both the CSIRO-PML and GLEAM products, while the MOD16 showed only half of that improvement. In other basins, the variability in the temporal pattern of degree correlations in time was still considerable and thus hindered any clear differentiation between the evaporation products. Even so, it was found that a constant lag of two months provided a better fit compared to other alternatives, including a zero lag. . Regardless of this finding, from a product assessment perspective, no significant or persistent advantage could be discriminated across any of the three evaporation products in terms of a sustained hydrological consistency with precipitation and water storage anomaly data. The results of this analysis have implications in terms of the confidence that can be placed in independent retrievals of the hydrological cycle, raise questions on inter-product quality and highlight the need for additional techniques to evaluate large-scale products.

The inability of evaporation models to directly capture , 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 and reservoirs is likely a factor in the variability of hydrological consistency observed between these related components. into our large scale water cycle studies.

1 Introduction

Space-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- in space and timefrom a completely new perspective. Indeed, remote sensing offers a number of independent means with which to retrieve various components of the hydrological cycle (e.g.i.e. rainfall, soil moisture, evaporation, terrestrial storage). 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 of space--based remote sensing is how 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 and satellite-based observationsmeasurements (McCabe et al. 2006).

Terrestrial evaporation (E), comprising the sources of soil and canopy evaporation together with plant transpiration, plays a key role in the water cycle as a linking mechanism between the surface and the atmosphere (Mueller et al., 2011). Unlike microwave or radiative emissions from the surface or atmosphere, which 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 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). When ground-based flux observations are available, thesethey can be used for calibration and evaluation;, but large-scale assessment is inevitably constrained by the lack of distributed and representative in situ networks to comprehensively assess simulations (Jana et al., 2016) ,as well as the inherent uncertainty ofassociated with these observations. Some recent evaluation efforts have sought to estimate the uncertainty of satellite-based models of evaporation, as well as those from land surface model and reanalysis data, in terms of the variance amongst the products (Mueller et al., 2011; Jimenez et al., 2011; Long et al., 2014). These and related attempts have shown that no single evaporation product consistently outperforms any other, whether applied at local (Ershadi et al., 2014) or global scales (Miralles et al., 2016). Considering this issue of spatial mismatch and model variability, it seems inappropriate to assess these large-scale products via direct comparison to in situ data alone. Moreover, the quality of any satellite-based product should not be judged solely on its agreement with potentially unrepresentative point-scale approaches. Central 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 ground- and satellite--based observations (McCabe et al. 2006). To compensate for this, it is important that a range of methods be used to evaluate the large-scale implementation of theseevaporation models.

ToA 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 (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. this end, A a limited number of studies have sought evaluated to quantify large-scale the water budgets closure of large basins across different regions of the world using either satellite observations alone (Sheffield et al., 2009) or via through 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 objectivesome of these studies (of Sheffield et al., (2009; Gao et al., 2010) was to evaluate the water budget closure (by comparing the residual of the water budget (i.e. inferred runoff) with measured streamflow values),runoff, the remainingother studies mostly aim to provide merged or observation /constrained estimates of the water cycle components, and anywith estimates of uncertainty is given in terms of the variability among the products (e.g. Long et al., 2014). The results of these studies have generally illustrated large water budget closure errors, focusing on the temporal scale and invoking the use of a hydrological model to guide analysis or force closure, rather than being solely observation driven assessments. Observation-only studies are important, as they provide an unbiased perspective not just on hydrological closure, but also allow for a first-order examination of the underlying agreement between component variables. However, rather than just comparing the uncertainties between the evaporation products and the other hydrological components (which are poorly defined), When considering these earlier contributions, there is still still rather than comparing the uncertainties between the evaporation products and the other hydrological components (which are poorly defined), remains a need for alternative assessment techniques that explore the connection between the hydrological variables at both temporal and spatial scales. One approach to determine this is to evaluate the hydrological consistency between observed products (McCabe et al., 2008). The term hydrological consistency refers to the spatial and temporal match that should inherently exist between independent observations of hydrological states and fluxes, based upon physical considerations. It is a concept that encompasses the expectation of water cycle behavior and mass balance: that is, changes in one term should be reflected in related variables, both spatially and temporally. For instance, a rainfall event should result in an observable change in soil water storage and a consequent increase in evaporative flux, which in turn should reduce the available soil moisture. This relatively simple concept has been explored in the past, including in efforts to improve precipitation events by employing cloud detection methodologies (Milewski et al., 2009); using soil moisture changes to infer precipitation amounts (Brocca et al., 2014); examining the connection between soil moisture state and changes in atmospheric variables such as humidity and sensible heat flux (McCabe et al., 2008); as well as in assessments of land–
-atmosphere coupling between observations and reanalysis data (Ferguson and Wood, 2010).

In considering these earlier contributions, there remains a need toto determine whether the goal basic idea of hydrological consistency can be realistically achieved extended to explore the agreement between using independent global-scale currently available satellite-based hydrological products. In To examineing this question, it makes sense to focus on regions catchments with that have relatively simple water budgetshydrological interactions, as they represent natural laboratories within which the evaluation of large-scale products and the concept of hydrological consistency could be reasonably undertaken. For example, Wang et al. (2014) evaluated the level of agreement between three satellite-based hydrological cycle variables over arid regions in Australia, where surface and subsurface runoff were minimal. Given a sufficiently low runoff component, a lack of snow accumulation, and a relatively strong coupling of precipitation and evaporation components, arid and semi-arid environments represent potential candidates within which to undertake such process assessments. Recognizing the need to embrace a more holistic evaluation strategy, this study seeks to explore the hydrological consistency within a number of basins where hydrological processes are relatively simple, i.e. given the conditions described above. Our analysis constitutes a framework for assessing the utility of hydrological consistency to evaluate remotely sensed hydrological products, with a focus here being placed on an assessment of recently developed global evaporation products (McCabe et al., 2016; Miralles et al., 2016). We undertake our analysis over four large river basins within arid and semi-arid environments distributed across the globe: the Colorado River basin in North America, the Niger River basin in Africa, the Aral Sea basin in Asia, and the Lake Eyre basin in Australia. The rationale behind the selection of these study basins and the potential deviations from the ideal conditions of a relatively simple hydrological system are explained in Section 2. , 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.

In compiling datasets with which to evaluate and differentiate the three evaporation products, a number of product specific considerations needed to be accounted forIn 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 estimates, which comprise 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). As any continuous function on a sphere, the gravity field can be represented as an expansion in spherical harmonics, which form a complete set of basis functions in the sphere: similar to the way a Fourier series expansion uses sines and cosines as basis functions. Unlike precipitation and evaporation products (and most other hydrological remote sensing variables), it is problematic to directly compare spatial maps of GRACE water storage data with other spatially distributed hydrological variables (Tapley et al., 2004a), since GRACE data are usually filtered in the spectral domain. While scaling the GRACE data to account for differences due to filtering has been proposed as an alternative solution to this problem (Landerer and Swenson, 2012), it has recently been shown to affect results in some cases (Long et al., 2015), including arid regions. Given this restriction, we implement an alternative approach in which the precipitation and evaporation fields are transformed into spherical harmonics in order to remove the impact (and model dependence) of this scaling term. This approach allows for a more reasonable and equivalent intercomparison of hydrological variables, and represents a key novel aspect of this work. Further details describing this process are presented in Section 3. 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). 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 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 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.

The oOverall objective of this study, the purpose of this paper is to evaluate the hydrological consistency of threeindependent satellite-based evaporation products with remotely sensed retrievals of precipitation and terrestrial water storage across a selection of basins that exhibit relatively well-defined hydrological interactions. Throughout this analysis we aim to determine whether the hydrological consistency concept can expand the range of evaluation metrics used to assess large-scale hydrological data sets such as evaporation, and enable some differentiation amongstof relative product quality them to be made. 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 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 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 recently developed global-scale evaporation products. In the following paragraphs we briefly describe the sources and nature of the data used in this contribution.

2.1 GRACE water storage anomalies

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^{-1} has been previously proposed (Wahr and Velicogna, 2006). Water storage anomalies (2003-2011) were derived fromcomputed using GRACE (release 05) monthly spherical harmonic coefficients representing the gravity field, processed at the University of Texas Center for Space Research (UTCSR). The gravity field is usually described in terms of the geoid,: an equipotential surface that is defined to correspond to the mean sea level over the oceans (Swenson and Wahr, 2002). The geoid is usually approximated as a linear combination of spherical harmonics, given that these represent solutions to the Laplace equation that describes the relation between the gravitational potential and the geoid. The approximation is of the form:

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

where $\tilde{P}_{lm}(\cos\theta)$ are the normalized associated Legendre functions, θ corresponds to the colatitude (the complementary angle to the latitude), ϕ to longitude, C_{lm} and S_{lm} are the spherical harmonic coefficients of degree l and order m, and l_{max} is the truncation degree. The total number of coefficients is given by $((l_{max}+1)^2 + l_{max}+1)/2$, while the resolution (the scale of the smallest feature of the gravity field that can be resolved using l_{max} coefficients) is approximately $\pi a/l_{max}$ (where a is the Earth's radius). The data product used in this study (processed by UTCSR) contains coefficients up to l_{max} = 60, i.e. a total number of 1891 coefficients, with an approximate resolution of 333 km. The full description of the process to transform the gravity field anomalies into water storage anomalies is described ~~in detail~~ in Wahr (1998) and Swenson and Wahr (2002).

GRACE data contains two types of errors (correlated and random) that need to be filtered before translating the data into water storage anomalies. Correlated errors are known to contaminate the signal in the form of north-south oriented stripes. A "de-striping" filter was applied to the coefficients (Swenson and Wahr, 2006; Duan et al, 2009) in order to remove this source of error. An isotropic filter (Gaussian filter with radius of 300 km) was then used to remove random errors (Swenson and Wahr, 2002). Furthermore, it is a usual practice to replace the ~~The~~ degree 2 coefficients with a more reliable estimate from a low-degree model of the gravity field calculated using ~~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~~filters ha~~s~~ves 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 surface model (LSM). This method ~~is~~ has been used to obtain scaling factors for GRACE data in order to restore the signal (it has been observed that the filters typically reduce the signal) before using the GRACE data with other hydrological variables ~~(~~(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~~ element of uncertainty in our study, which is focused on arid regions, we instead transformed the other water cycle components (i.e. evaporation and precipitation) into spherical harmonics, using an approximation similar to equation 1. The effect of the filters is therefore incorporated directly into the ~~study by applying the filters to the various~~other hydrological components in 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., 2011a). 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 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 briefly 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 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 lLeaf aArea iIndex (LAI) and Normalized Difference Vegetation Index (NDVI)), as measured by the MODerate resolution Imaging 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)), a three--source scheme used for terrestrial land flux estimation. 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 adjustments includeincorporated into MOD16 include: 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 MOD16 -product includes comprises transpiration, and evaporation from the soil and wet canopy, as well as total evaporation calculated as the sum of these three components. Each component is weighted-based on the fractional vegetation cover, relative surface wetness and available energy. Inputs to the model include net radiation (R_n), air temperature and humidity, as well as LAI and vegetation phenology. Importantly, it does not require wind speed or soil moisture data, making it a relatively parsimonious model in terms of input requirements. In this study, we used the actual evaporation (AET) product from MOD16 (Mu et al., 2011) with 8-day temporal resolution and 1 km resolution in the sinusoidal projection. The product was reprojected onto a 0.05° regular grid using the MODIS Reprojection Tool (MRT) before transformation into spherical harmonics, as described in Section 3.1. Further details on the modeling basis behind the MOD16 product can be found in Mu et al. (2013), Ershadi et al., (2014) and Michel et al. (2016). Potential evaporation (calculated using a Priestley-Taylor based formulation) is also included to monitor environmental water stresses and droughts (Mu et al., 2011).

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 two-source Penman–-Monteith–-Leuning (PML) model. An important new feature of the PML approach was a biophysical algorithm for the calculation of the canopy conductancesurface resistance, which was previously calculated as LAI multiplied by a constant c_L (Cleugh et al., 2007). The new parameterization of the surface resistance in the PML model was optimized using data from 15 globally distributed flux station sites, andwith two key parameters were identified: the maximum stomatal conductance ($g_{sx}$the conductance is the inverse of resistance) and the ratio of actual to potential evaporation at the soil surface. Zhang et al. (2010) developed a method to further optimize the spatial variability of these two parameters (i.e. at each grid pixel) using gridded meteorological data and a simple Budyko-curve hydrometeorological model developed by Fu (1981) that includes precipitation and available energy as inputs. Mean annual evaporation for each grid pixel is calculated using the Fu model and gridded meteorological data. Then, tThe value of g_{sx} (i.e., the maximum stomatal conductance) is optimized using a non-linear least square regression-based on the difference between the PML and the Fu model. Interestingly, the Fu model is calibrated by comparing the output evaporation with the residual of precipitation and runoff, i.e. by assuming negligible annual water storage changes and groundwater inflow and outflow. –Zhang et al. (2012) used this approach to develop a global gridded terrestrial evaporation product (hereafter referred to as CSIRO-PML; Zhang, 2014, personal communication) with a 0.25° resolution (in this study, we used the actual evaporation product). They used gridded meteorological data from diverse sources:, including vapor pressure and temperature from the Climate Research Unit (New et al., 2000), LAI and land cover type from Boston University (Ganguly et al., 2008), precipitation from the Global Precipitation Climatology Centre (GPCC, version 4; Rudolf and Schneider 2004), and radiation data from the Global Energy and Water Cycle Exchanges (GEWEX) Surface Radiation Budget (Gupta et al., 2006).

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., 2011a) 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) as a first step. GLEAM then employs the Priestley−Taylor equation to calculate the potential evaporation of bare soil and vegetation components (both short and tall canopy), with values constrained to actual evaporation via application of a stress factor. The stress factor is calculated using vegetation optical depth from a combination of different satellite passive microwave observations using the Land Parameter Retrieval Model (Liu et al. 2013). GLEAM also has the capacity to explicitly calculate sublimation of snow covered surfaces (Globsnow snow water equivalent product, Takala et al., 2011) as well as open water evaporation. . Next, sSatellite observations of surface soil moisture are can be assimilated using a Kalman filter assimilation approach to estimate the moisture profile over several soil layers. Here we employ version 2A of GLEAM (Miralles, 2014, personal communication), which uses a combination of satellite, ground and reanalysis input data. Precipitation is obtained from the Climate Prediction Center Unified data set, consisting of data from over 30,000 stations (CPC-Unified, Joyce et al., 2004). The radiation product used in this version of GLEAM is the European Center for Medium-Range Weather Forecasts (ECMWF) ERA-Interim meteorological reanalysis product (Dee et al., 2011). In this version of GLEAM, surface soil moisture data from the Water Cycle observation Multi-mission Strategy Climate Change Initiative (WACMOS-CCI) merged product (from a combination of several passive and active microwave products) is assimilated (Liu et al., 2012), while air temperature is derived from both the International

Satellite Cloud Climatology Project (ISCCP) and the Atmospheric Infrared Sounder (AIRS) (Rossow and Dueñas, 2011). Further details of the model can be found in Miralles et al. (2010), Miralles et al. (2011a) and Miralles et al. (2011b).

~~. 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 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 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 the best available sources of precipitation data and has been previously used in soil moisture- and evaporation-based analyses (Crow, 2007; Miralles et al., 2011a). 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

While runoff data was not used explicitly in the consistency analysis presented in this manuscript, ~~Although not considered in the evaluation of hydrological consistency,~~simulated runoff data was~~is~~ compared to precipitation and evaporation observations~~data~~ in order to evaluate ~~whether~~ the assumption of a relatively simple ~~water budget~~hydrological system ~~is valid~~ in the study ~~regions~~basins. 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. Although these values were not constrained with ground estimates and thus may contain biases, as noted, runoff values were only used to provide an assessment of runoff against the observed precipitation and evaporation data. The version of the product used in this study (GLDAS-2.0) is forced with meteorological data from the Princeton University forcing data set (Sheffield et al., 2006) and is available at 1° ~~-degree~~ resolution from 1948 to 2010.

2.5 Selection of s~~S~~tudy ~~regions~~basins

The study basins~~regions~~ were ~~chosen~~ targeted ~~targeted~~ primarily~~mainly~~ on their climate classification, with river basins in~~that covered~~ regions with a predominantly arid or semi-arid climate preferentially selected. This criterion was established in order to seek a relatively simple hydrological system (i.e. constrain the range of possible hydrological interactions), thereby maximizing the conditions under which hydrological consistency between evaporation and precipitation and water storage changes might be achieved. A ~~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), was used to identify arid~~-~~ and semi-arid regions. ~~The climate criteria was to 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) and~~,~~ derived from flow direction data of the HYDRO1k Elevation Derivative Database, developed at the U.S. Geological Survey (USGS). A threshold of 50% areal extent containing any of the arid Köppen climates (BWk, BWh, BSk or BSh) was used to select potential basins. Secondary ~~considerations~~ criteria for ~~selecting~~ basin selection~~s~~ from the GRDC data set ~~were based~~focused on size, geographical ~~location~~ distribution and amplitude and trends in the water storage variations. In terms of the size of the basin, a smaller size would more likely satisfy the assumption of a relatively simple hydrological system. However, due to the coarse resolution of GRACE data (see section 2.1), this requirement had to be compromised. Given these considerations, ~~We selected~~ four basins were selected as focus regions of study: the Colorado R~~R~~iver basin (CRB) in North America, the Niger River basin (NRB) in Africa, the Aral Sea basin (ASB) in Asia and the Lake Eyre basin (LEB) 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~~ establish the extent to which a major assumption of the study ~~(y,~~ i.e. a simple water budget~~),~~ is met 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 R~~R~~iver basin ~~(CRB)~~ and the Aral S~~S~~ea 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~~Likewise, the Niger R~~R~~iver basin ~~(NRB)~~ also has a runoff component that is ~~similar in~~ close in magnitude to~~as important as~~ evaporation, but we assume that it will not affect the spatial distribution of water storage anomalies. ~~Finally,~~I~~ i~~n the Lake E~~E~~yre basin ~~(LEB)~~ we expect that the limited and sporadic runoff component will not have a significant effect on the hydrological consistency analysis undertaken here.

Even though these four basins were preselected based upon their location within dryland systems (Wang et al., 2012), they reflect a range of trends in water storage and precipitation. For example, the Colorado River basin experienced intervals of wet and dry periods, while the Niger River basin showed a small but steady increase in water storage with a clear ~~,~~seasonal variability in both water storage and precipitation. Meanwhile, the Aral Sea basin experienced a significant loss of water during the study period (-8.2 mm.yr^{-1}), in line with the historical depletion of this inland sea in response to increased agricultural productivity (Zmijewski and Becker, 2014). Finally, tThe Lake Eyre basin showed a marked increase in precipitation during the end of the study period (2009-2011), with a corresponding increase in water storage during the following years, reflecting the larger scale hydrometeorological conditions affecting that region (Boening et al., 2012).

3 Methodology

In order to provide a meaningful spatial evaluation of the hydrological consistency between the data sets (i.e. at sub-basin scale) and to ensure that a fair comparison between GRACE data and satellite products could be undertaken, the analysis was carried out in spherical harmonics. Furthermore, theThe effects of the de-striping filter (see Section 2.1) are incorporated into the analysis directly 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 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 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):

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

where $\tilde{P}_{lm}(\cos\theta)$ is the normalized associated Legendre functions, θ corresponds to the colatitude (the complementary angle 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 ≈ 330 km.

The spherical harmonic analysis refers to the process of solving equation 1 for a set of coefficients C_{lm} and S_{lm} up to an approximation of degree l_{max}. Several computational packages are available to perform this type of analysis. In this study,Here 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.

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~~product. Next, we applied the de-striping and Gaussian filters to account for the effect that these have in GRACE TWSA data (see section 2.1).

3.2 Regional spherical harmonic analysis

In the analysis so far, t~~T~~he computed spherical harmonic coefficients ~~so far~~ are global (e.g. Figure 3). In order to evaluate the hydrological consistency of the study regions (Figure 1), the data ~~needed~~ needs to be masked for the particular ~~study regions~~basins. 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~~ the 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 averaging kernel following Eq. (2):

$$f^b(\theta, \phi) = f^g(\theta, \phi)\vartheta(\theta, \phi), \qquad (2)$$

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):

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

$$Q^{jm}_{j_1 m_1 j_2 m_2} = \sqrt{\frac{(2j_1+1)(2j_2+1)}{4\pi(2j+1)}}\, C^{j0}_{j_1 0 j_2 0} C^{jm}_{j_1 m_1 j_2 m_2}, \qquad (4)$$

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

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

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

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

The degree correlation measure is computed for every degree (l), and therefore we can in principle evaluate the hydrological 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~~ basin in this study is the Colorado ~~River~~ 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?~~ An ~~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~~ 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 River and the Lake Eyre basins, immediately ~~constraining~~ negating any type of hydrological consistency analysis. In the Niger ~~river~~ 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 Colorado River basin~~RB 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, even from ~~even~~ ~~just~~ a qualitative perspective, there are clear challenges in developing a hydrological ~~closure~~ consistency approach over these comparatively "simple" basins. Indeed, this has been demonstrated in other studies using either satellite data alone, or a combination of satellite and ground data. While it is not the intent of the~~is~~ current work to explore the error characterization of these different evaporation models based on hydrological closure, the techniques being used to evaluate product consistency ~~do~~ should provide some insight into retrieval quality: at least relative to the other hydrological products (precipitation and gravity-based water storage changes) that the evaporation is 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~~In this section, we examine~~d~~ the spatial and temporal patterns of the degree correlations between water storage variations (TWSA) and P-E anomalies. Figure~~s~~ 4 to Figure ~~-~~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 e.g. ~~. For example,~~ do the degree correlations behave differently during wet or dry periods, or when storage changes ~~can be attributed to either~~are ~~caused~~driven by 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 R~~r~~iver b~~b~~asin ~~(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)~~The start of the study period (2003) coincided with the end of an intense multi-year drought in the Colorado River basin (Scanlon et al., 2015). During the wet period of 2004-2005 (Figures 2), ~~the CRB~~~~this~~the ~~region~~ basin showed a corresponding increase in TWSA (see 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~~2005), 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, since~~:~~ a large portion of the runoff comes from snowmelt originating in the upper Colorado R~~R~~iver b~~B~~asin (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.~~,~~ H~~h~~owever, 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~~ River ~~basin~~ basin ~~(NRB)~~

The TWSA in this ~~region~~ basin was characterized by an overall steady increase (5.79 mm.yr^{-1}) ~~and a~~with clear seasonal variability (see Figure 5). Over the study period, p~~P~~recipitation ~~peaked~~ peaks between July and September, while TWSA ~~peaked~~ peaks between September and November. Ahmed et al. (2014) attributed the observed 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 (2003-2004), ~~the region~~this basin 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 Colorado River basin~~RB~~ 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 observations of TWSA and P-E. Interestingly, there ~~also~~ seems to be less inter-degree variability compared to the other ~~regions~~study basins. This may be related to the simpler water budget in this ~~region~~ basin compared to that of the Colorado ~~r~~River and Aral Sea basin~~RB and~~s ~~ASB~~, but requires further investigation.

4.2.3 Aral Sea b~~b~~asin ~~(ASB)~~

Th~~e Aral Sea basin is an~~ is endorheic basin ~~that has experienced~~continued a historical trend of water loss ~~a significant loss of water~~ during the study period, most likely caused by agricultural activities (~~(-8.2 mm.yr^{-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 Colorado River ~~R~~basin~~B region~~: that is, a weak connection between high r_l values and increasing or decreasing TWSA before reaching the local maxima or minima, respectively (see Figure 6). 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~~ basin during several months. Although the Aral Sea ~~S~~basin~~B~~ is predominantly 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 inputs, the evaporation products may represent these interinter-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 regionsbasins), as well as in the slight differences in degree correlation between the evaporation products.. Differences in degree correlation due to the use of the three evaporation products were minimal i.e. no single evaporation product resulted in a significantly higher (or lower) hydrological consistency with precipitation and water storage anomalies.

4.2.4 Lake Eyre bbasin (LEB)

Another endorheic basin examined here was the Lake Eyre basin, which experienced a marked increase in precipitation (Figure 2) during the rainy seasons in of 2009-2011, resulting in an increase in water storage anomalies of about 40 mm/year.yr^{-1} (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 (see Figure 7). 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 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 r_l, 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 generalOverall, 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. More importantly, none of the evaporation products showed a significant (and persistent) advantage in terms of hydrological consistency over the others. Wang et al. (2014) also studied the hydrological consistency of satellite products (TRMM-based P, MOD16-based E and GRACE TWSA) over the LEBthis basin, as well as other predominantly arid regions of 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 reachup 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, during which as water accumulates in the soil and travels from different source 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 from amongst thethose basins studied here is in the Niger River basinRB (Figure 5), where a lag of two months is evident throughout the study period. In other regions, however, such as the Colorado River basin ~~RB~~ and Lake Eyre basin~~B~~, 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 geomorphological characteristics (Ahmed et al., 2011; Wang et al., 2014). Because of their large extent and geographical features, the Colorado river and Aral Sea basins include regions where snow storage plays an important role as delayed sources of runoff. Other potential sources of delayed flow, not limited to these two basins, include groundwater flow and surface water flow. The combination of these slowly varying components of streamflow is defined as baseflow (Beck et al., 2013).

To examine this temporal component, at least in a simplified manner,~~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 Niger River ~~R~~basin~~B~~ 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 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). Beck et al. (2013) developed global estimates of the Base Flow Index (BFI):~~,~~ a measure of the ratio of the long-term baseflow to the total runoff, using a large global data set of runoff and a regionalization procedure to transfer these and other characteristics of runoff from gauged to ungauged basins. ~~Because~~Since we did not model any of the physical processes contributing to baseflow, the BFI ~~could~~was examined to assist in~~ help~~ ~~-~~explaining part of the delay in observed water storage changes relative to the P-E term (~~,~~although not dynamically since ~~(~~the index is a long-term average in time)~~)~~. The spatial average of the BFI in the four study basins is, not surprisingly, within the same range: between 0.4 and 0.6. This is not unexpected, as various climate characteristics were used as predictors of BFI. ~~Furthermore~~Indeed, the fact that they are similar is in agreement with our finding of similar GRACE lag times among the study basins. Further investigation is required to determine the nature of the elements affecting the lag in water storage, not limited to those found in baseflow.

Figure 8 presents a statistical summary of the mean degree correlation values over the study period comparing the original 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~~ Niger ~~r~~River basin 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 outcome was similar irrespective of the evaporation product used. For the ~~CRB~~ Colorado ~~r~~River 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~~ Aral ~~S~~sea basin~~region~~ also benefited from an imposed lag in GRACE data, although there remained considerable variability in the results. The ~~LEB~~ Lake Eyre basin~~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, t~~The development of methods and sensors to retrieve the various components of the water cycle has for the most part~~largely~~ been undertaken independently of ~~other (~~interrelated~~) variables~~ processes (see McCabe et al., 2008 and Brocca et al., 2014 for some examples of complementary retrieval). Large-scale retrievals of hydrological variables such as evaporation, soil moisture and rainfall products do not come with well-defined accuracy metrics, let alone uncertainty bounds. ~~The~~This lack of any well-defined error structure associated with individual products complicates the task of product assessment. As such, the question of how to evaluate ~~these~~large-scale datasets remains an outstanding one. This is especially important in the context of global-scale products. While a number of global evaporation (and precipitation) evaluation papers have been published, none seek to identify consistency with related hydrological variables, and focus instead on comparisons against traditional point-scale or tower-based techniques (McCabe et al., 2016; Michel et al., 2016). Given the spatial mismatch between ground observations (and the lack of continuous large-scale coverage of in situ data in remote regions), it is inappropriate to evaluate these large-scale products in such a manner. Determining whether individual products are at the least consistent with each other (i.e.~~,~~ they reflect hydrological expectation) is a needed first step in product assessment. The motivation behind this study was to take a step back and determine whether a first order hydrological assessment could be achieved. Rather than comparing the uncertainties between the evaporation products and the other hydrological components (which are poorly defined), we attempt to distinguish between the different evaporation products relative to their consistency with precipitation and terrestrial water storage. That is, are observed changes or patterns in the evaporation datasets reflected in these other hydrological variables? We explore this approach precisely because of the challenges in quantifying uncertainty based upon traditional in situ methods. As is discussed below, the challenge on how to do this remains, raising some important questions on both product quality and also the techniques we use to evaluate global products.

5.1 Challenges to implementing hydrological consistency

It is~~, 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 inter-product agreement between hydrological components should exist~~might~~

~~be found~~. ~~For this reason~~To explore this idea, our~~the~~ study ~~assumed~~ focused on basins where a~~a~~ simpler water budget, consisting of water storage anomalies as a function of precipitation and evaporative fluxes, might be expected to predominate. The aim was to limit potentially complicating variables such as snow, vegetation changes, large precipitation and streamflow contributions and other hydrological processes. The assumption was that arid~~-~~ and semi-arid regions would best fit this profile. ~~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 geographical~~ly~~ distribution~~ed~~ ~~selection~~ of basins that could satisfy~~fit into~~ this simplified water budget assumption ~~is somewhat limited~~was difficult. Restrictions related to basin size affect the study in two conflicting ways. On the one hand, a large basin will inevitably present complications related to heterogeneity (including in climate zones, as was the case for the Colorado River basin and Aral S~~s~~ea basin) and also be more likely to contain areas affected by anthropogenic activities, such as irrigation, land cover changes, building of dams and reservoirs, etc. On the other hand, a small catchment size would be more difficult to evaluate with this consistency approach, given the coarse resolution of (most) of the global products used here, but especially the GRACE data. The spatial resolution of GRACE data is further limited by the use of filters to remove errors. Considering these restrictions, a compromise in the selection of study basins was required to allow for at least a narrow range of length scales (500-800 km) to be evaluated.

In the end, our study consisted of four major globally distributed river basins, including two endorheic systems. Although having mostly an arid climate in terms of Köppen classification, both the Colorado ~~r~~River and Aral Sea basins include regions with the presence of snow and snowmelt-dominated runoff. While snow storage itself is not a problem, since GRACE detects changes in storage irrespective of their nature (snow, groundwater, soil moisture, etc), snowmelt may contribute to delayed changes in storage that can affect gravity results. As such, the inclusion of these basins was considered important in order to test the hydrological consistency concept in regions that deviated from the ideal assumption. The influence that snowmelt and other potential sources of lag in the system have is poorly defined and forms part of the motivation to explore the inclusion of a lag response in the GRACE data (see Section 4.3).

Apart from the issues of spatial scale, the use of satellite-based hydrological data presents additional challenges and sources of uncertainty to a consistency-based assessment. For instance, because GRACE data is smoothed to remove errors in short-scale terms (i.e.~~,~~ truncation of the spherical harmonic coefficients), the gravity signal contains contamination from outside of the studied basins (leakage) and is a potential source of uncertainty in areas neighbouring high amplitude signals (particularly if they are out of phase~~single~~ ~~signal~~with the study basin) and the ocean. Although the LSM-based scaling factor, which is static in time, has been used to correct for bias (e.g. signal reduction) and leakage contamination, dynamic changes in water storage trends outside the basin might still contaminate the signal (Long et al., 2015)~~-~~. In addition, the temporal lag in terrestrial storage response as documented in previous studies (Rieser et al., 2010; Ahmed et al., 2011; Wang et al., 2014; Ayman and Jin, 2016) and observed in our analysis, ~~was~~represents an important source of potential error (see Sections 4.3 and 5.2). Product errors are also evident in the precipitation and evaporation data sets. Global rainfall retrievals have well recognised limitations, including the detection of both high and low intensity events (Hou et al., 2014), the discrimination of cloud clear and cloud precipitation scenes, as well as the sensitivity to parameters in the forward model of radiative transfer over different sensors (Stephens and Kummerow, 2007). In terms of evaporation, uncertainties related to algorithm choice, input data variability and process parameterizations all complicate the accurate estimation of terrestrial fluxes (Ershadi et al. 2015). Determining whether or not and understanding how much these product sources of uncertainty affect hydrological consistency studies remains an area requiring further investigation.

5.2 Temporal lag in terrestrial storage response

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

In exploring the relationship between GRACE water storage changes and precipitation and evaporation data, it was evident that water storage anomalies peaked at a significantly later time than the corresponding P-E values. One possibility for this apparent lag in ~~that has been explored in studies using~~ GRACE data ~~(Rieser et al., 2010; Ahmed et al., 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 (a rough estimate of GRACE accuracy averaged over the entire Earth is 20 mm.month^{-1}; Wahr and Velicogna, 2006),~~,~~ ~~and therefore~~ the corresponding mass is not detected until a sufficient amount has accumulated within the catchment via natural drainage processes (Rieser et al., 2010; Ahmed et al., 2011)~~.~~. 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 whole study period. Doing t~~T~~his ~~seemed to~~ improved 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 ~~r~~R~~R~~iver 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 R~~r~~~~R~~iver basin and Aral Sea basin, where changing trends in the seasonal patterns of precipitation make made 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 effect on the degree correlation. Further understanding the implications and physical rationale behind the attribution of this lag is required.

5.3 Discriminating between satellite evaporation products

A secondary objectiveOne aspect of this work was to determine explore whether differences in available evaporation products could affect the results of the consistency analysis i.e. could we identify was there better agreement between water storage anomalies and P-E with in anyone particular evaporation product? While tThe analysed products used in the study covered a wide range of resolutions (0.05°–0.25°), although the effective resolution in the analysis is was ultimately determined by the truncation degree (l_{max}) of the spherical harmonic transformation. Even after accounting for this, absolute differences were evident from a qualitative basin-scale analysis (Figure 2). Overall, rResults 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 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. Whether these differences in model parameterization are the sole cause of the apparent underestimation by MOD16 remains to be investigated.

BecauseAs the focus of the study was to discriminate between evaporation products, the question on whether the choice of precipitation product affects the hydrological consistency analysis remainswas somewhat beyond the scope of this work. However, a first orderpreliminary analysis was undertaken by replacing the GPCP precipitation fielddata with another data set and reproducing the analysis. To do this, we processed the Precipitation Estimation from Remote Sensing Information using Artificial Neural Network (PERSIANN) product, which uses an aArtificial neural network to approximate spatiotemporal non-linear relationships between physical variables and remotely sensed signals (Hsu et al., 1997). PERSIANN uses data from the long wave infrared imager onboard the Geostationary Operational Environmental Satellite (GOES) as well as from the Tropical Measurement Mission (TRMM) microwave imager (TMI). As shown in Figure S1, Tthe results of this new analysis did not showcasereveal any significant difference thanwhen compared to those based on with the GPCP analysis. Figure S1 shows the average degree correlation statistics per study region and evaporation product, with and without the inclusion of a lag in GRACE data.

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

Evaluating global evaporation products remains an outstanding challenge. The purpose of implementing a hydrological consistency approach was to explore the evaluation of evaporative fluxes by comparing the spatial patterns between precipitation and changes in water storage. If such an approach could be shown to perform well in a relatively simple hydrological system, the potential for broader-scale application in regions with more complex behaviour would be the next logical step. However, the study showed that even in these relatively simple basins, it was not possible to demonstrate a consistent hydrological agreement between ~~current~~ independent observations. Improvements in satellite-based evaporation products ~~will~~are likely to be delivered 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~~anticipated to provide more accurate estimates of the water storage anomalies. Although products will inevitably improve (or continue to improve in an individual manner), it will ~~be~~remain important to evaluate them in a holistic manner. ~~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 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.

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 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 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.

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. 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 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, a key ~~one of the~~ motivation behind~~ng elements of~~ this study was to examine the capacity of independent observations ~~(or estimates)~~ of the water cycle to ~~achieve~~ reflect some form of hydrological ~~closure~~consistency. To do ~~this~~so, the study focused on regions where it would be most expected to achieve ~~hydrological consistency~~this: ~~that is,~~ in arid and semi-arid regions with a simplified water budget, consisting ~~(ideally~~primarily ~~)~~ of ~~only~~ precipitation and evaporation. Unfortunately, w~~W~~e found that even in these simple environments, hydrological consistency was difficult to obtain~~ throughout the study studied period~~. While there are times and locations at which some consistency was observed, there are a 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 within individual products. ~~Furthermore, a~~Although there were significant and known differences in evaporation estimates, especially with the MOD16 product in the Colorado river and Aral S~~s~~ea basins, these differences did not seem to play a significant role in the evaluation of hydrological consistency.

While ~~Although~~ not providing a comprehensive tool for product evaluation, the approach did help to reveal some interesting spatial and temporal patterns between the studied hydrological variables.

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 R~~R~~iver 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 them up to the point where these were minimal, ~~then~~ with 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 quite 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 to account for delayed sources of water storage changes, 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 average degree correlation and variability of the results in time.

The lack of persistent agreement in ~~other~~ some of the studied basins ~~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 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 an element of a holistic approach to ~~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 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:10.1175/1525-7541(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 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, Geodesy and Geodynamics, doi:http://dx.doi.org/10.1016/j.geog.2016.03.002, 2016.

Beck, H. E., Dijk, A. I. J. M. van, Miralles, D. G., Jeu, R. A. M. de, Bruijnzeel, L. A. (Sampurno), McVicar, T. R. and Schellekens, J.: Global patterns in base flow index and recession based on streamflow observations from 3394 catchments, Water Resources Research, 49(12), 7843-7863, doi:10.1002/2013WR013918, 2013.

Boening, C., Willis, J. K., Landerer, F. W., Nerem, R. S. and Fasullo, J.: The 2011 la niña: So strong, the oceans fell, Geophysical Research Letters, 39(19), doi:10.1029/2012GL053055, 2012.

Brocca, L., Ciabatta, L., Massari, C., Moramarco, T., Hahn, S., Hasenauer, S., Kidd, R., Dorigo, W., Wagner, W. and Levizzani, V.: Soil as a natural rain gauge: Estimating global rainfall from satellite soil moisture data, Journal of Geophysical Research: Atmospheres, 119(9), 5128-5141, doi:10.1002/2014JD021489, 2014.

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.

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.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., Berg, L. van de, Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., Rosnay, P. de, Tavolato, C., Thépaut, J.-N. and Vitart, F.: The ERA-interim reanalysis: Configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137(656), 553-597, doi:10.1002/qj.828, 2011.

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.

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.

Famiglietti, J. S., Cazenave, A., Eicker, A., Reager, J. T., Rodell, M. and Velicogna, I.: Satellites provide the big picture, Science, 349(6249), 684â€“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.

Ferguson, C. R. and Wood, E. F.: Observed land-Atmosphere coupling from satellite remote sensing and reanalysis, Journal of Hydrometeorology, 12(6), 1221-1254, doi:10.1175/2011JHM1380.1, 2011.

Fu, B. P.: On the calculation of the evaporation from land surface (in Chinese), Sci. Atmos. Sin., 5, 23–31, 1981.

Gao, H., Tang, Q., Ferguson, C. R., Wood, E. F. and Lettenmaier, D. P.: Estimating the water budget of major US river basins via remote sensing, International Journal of Remote Sensing, 31(14), 3955-3978, doi:10.1080/01431161.2010.483488, 2010.

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.

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, doi:10.1175/1525-7541(2001)002<0036:GPAODD>2.0.CO;2, 2001.

Hsu, K.-l., Gao, X., Sorooshian, S. and Gupta, H. V.: Precipitation estimation from remotely sensed information using artificial neural networks, Journal of Applied Meteorology, 36(9), 1176-1190, 1997.

Jana, R. B., Ershadi, A. and McCabe, M. F.: Examining the relationship between intermediate scale soil moisture and terrestrial evaporation within a semi-arid grassland, Hydrology and Earth System Sciences Discussions, 2016, 1-27, doi:10.5194/hess-2016-186, 2016.

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.

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.

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.

Liu, Y. Y., Dijk, A. I. J. M. van, McCabe, M. F., Evans, J. P. and Jeu, R. A. M. de: Global vegetation biomass change (1988-2008) and attribution to environmental and human drivers, Global Ecology and Biogeography, 22(6), 692-705, doi:10.1111/geb.12024, 2013.

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.

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 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.

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 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., Gash, J. H., Holmes, T. R. H., Jeu, R. A. M. de and Dolman, A. J.: Global canopy interception from satellite observations, Journal of Geophysical Research: Atmospheres, 115(D16), doi:10.1029/2009JD013530, 2010.

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, 2011a.

Miralles, D. G., De Jeu, R. A. M., Gash, J. H., Holmes, T. R. H. and Dolman, A. J.: Magnitude and variability of land evaporation and its components at the global scale, Hydrology and Earth System Sciences, 15(3), 967-981, doi:10.5194/hess-15-967-2011, 2011b.

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.

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.

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., WÃ³jcik, 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.

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, 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, 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.

Rossow, W. B. and Dueñas, E. N.: The international satellite cloud climatology project (ISCCP) web site: An online resource for research, Bulletin of the American Meteorological Society, 85(2), 167-172, doi:10.1175/BAMS-85-2-167, 2004.

Sahoo, A. K., Pan, M., Troy, T. J., Vinukollu, R. K., Sheffield, J. and Wood, E. F.: Reconciling the global terrestrial water 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 Ccolorado river basin, Water Resources Research, 51(12), 9891-â€“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.

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 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, 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.

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.

Takala, M., K. Luojus, J. Pulliainen, C. Derksen, J. Lemmetyinen, J. Kärnä, J. Koskinen, and B. Bojkov.: Estimating Northern Hemisphere Snow Water Equivalent for Climate Research through Assimilation of Space-Borne Radiometer Data and Ground-Based Measurements, Remote Sensing of Environment. 115. 3517-3529. http://dx.doi.org/10.1016/j.rse.2011.08.014, 2011.

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.

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 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.

Wang, L., D'Odorico, P., Evans, J. P., Eldridge, D. J., McCabe, M. F., Caylor, K. K. and King, E. G.: Dryland ecohydrology and climate change: Critical issues and technical advances, Hydrology and Earth System Sciences, 16(8), 2585-2603, doi:10.5194/hess-16-2585-2012, 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 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~~ Aral S~~s~~ea 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. (2011a,b)
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). The original MOD16 product is available at 1 km resolution in the sinusoidal projection. In this study, the product was reprojected onto a 0.05° regular grid using the MODIS Reprojection Tool (MRT).

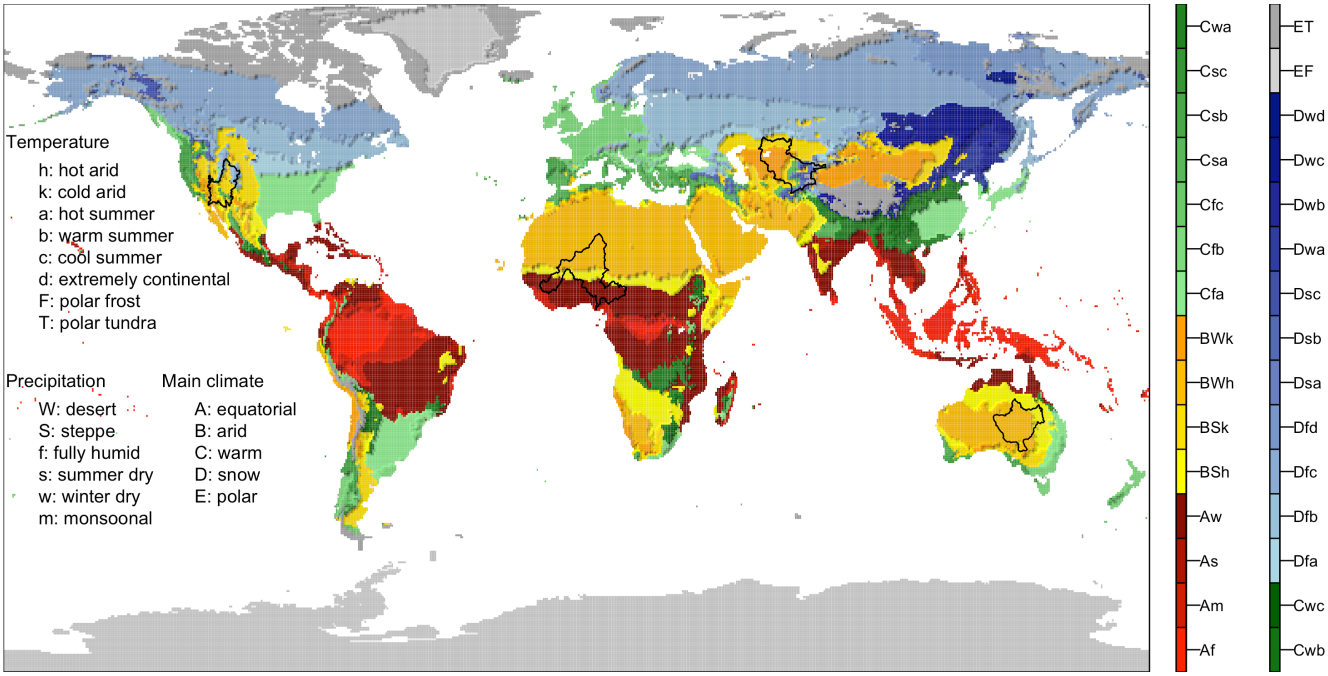

Figure 1: Selected study basins used within the analysis. Criteria for the selection of basins included: predominantly arid climate (more than 50% areal coverage with any of the arid Köppen climates: BWk, BWh, BSk or BSh), size, geographical location and amplitude and trends in the water storage variations.

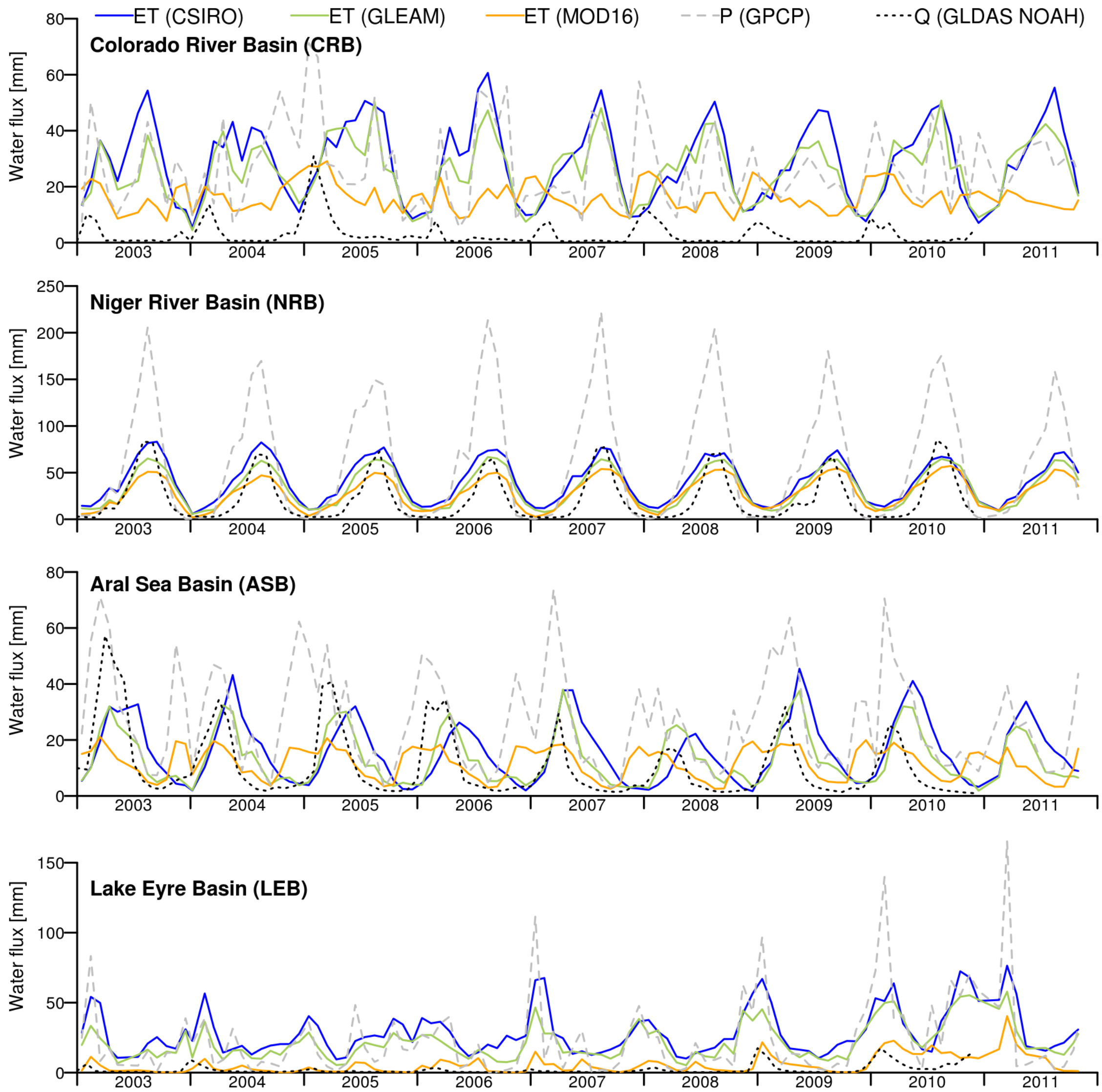

Figure 2. Average P, E and Q fluxes within the four study basins for the period 2003-2011. Three evaporation (E) data sets were used in this study, including CSIRO-PML, MOD16 and GLEAM (see Table 1 for details).

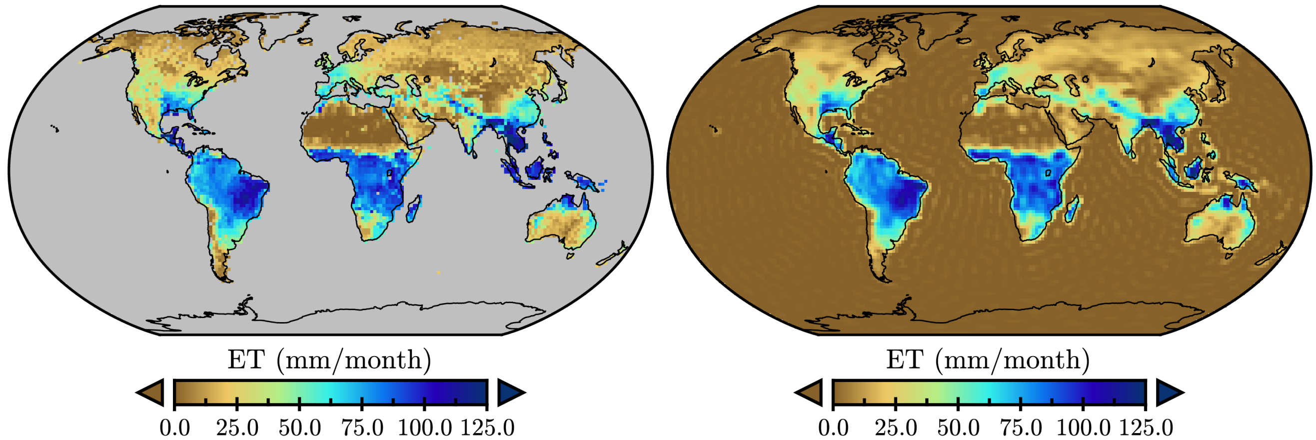

Figure 3. Left: CSIRO-PML monthly evaporation (E) for April 2003. Right: the same data set after spherical harmonic analysis and synthesis of the evaporation data. Missing data are set to zero. The data appears smoothed because it is an approximation (equation 1) limited by l_{max}. There are also other effects such as ringing that are visible in regions where the data is close to zero.

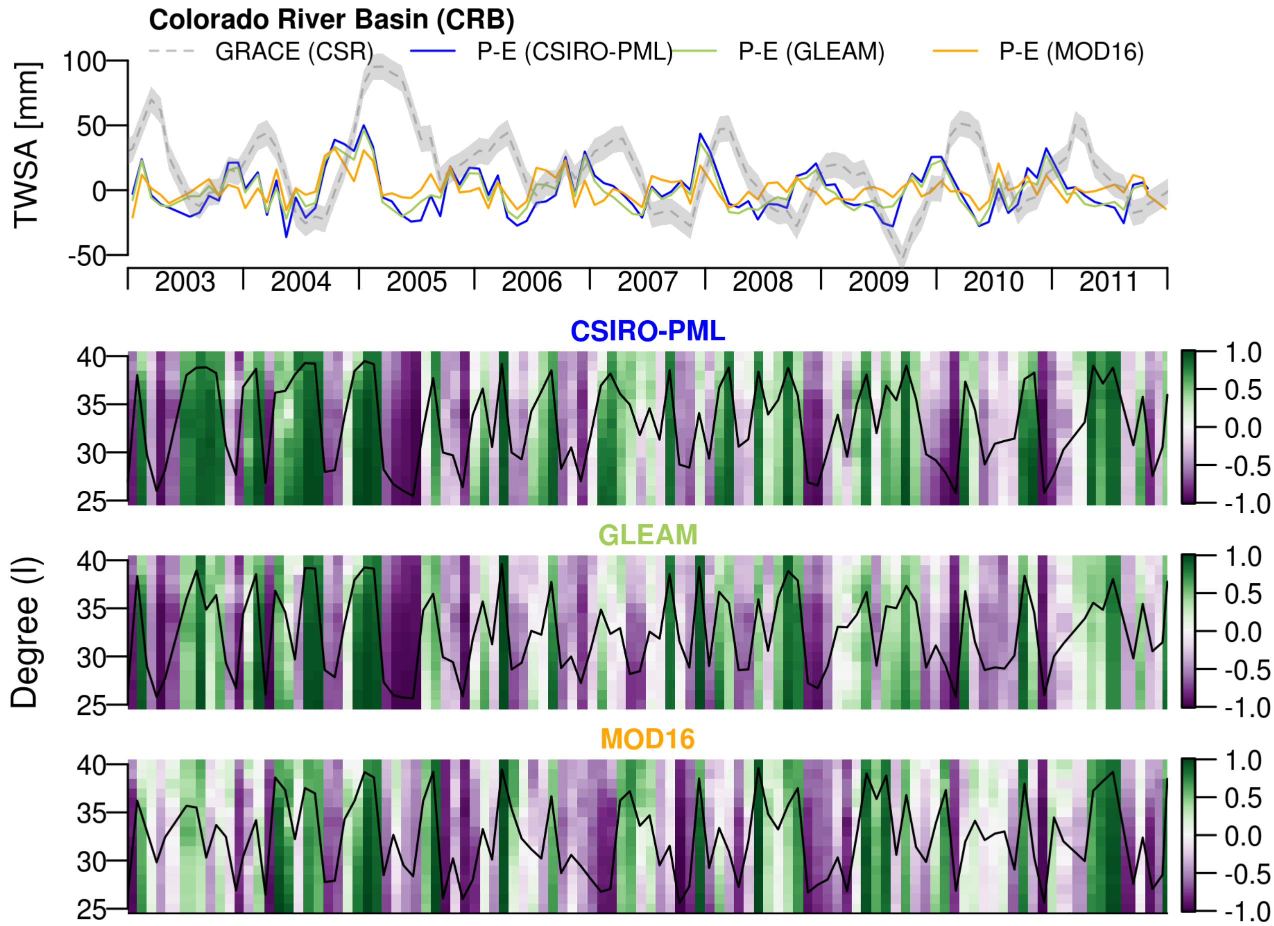

Figure 4. Top: anomalies of the terrestrial water storage (TWSA) observed by GRACE (with 20 mm uncertainty bounds) and P-E using three global evaporation products over the Colorado river basin. Below: varying degree correlation measure (r_l) with time and degree (from 25 to 40) using the three global evaporation products. The average r_l is shown as a time series (black line). The degree correlation measure can range from -1 to 1 as shown in the color scale on the secondary axis.

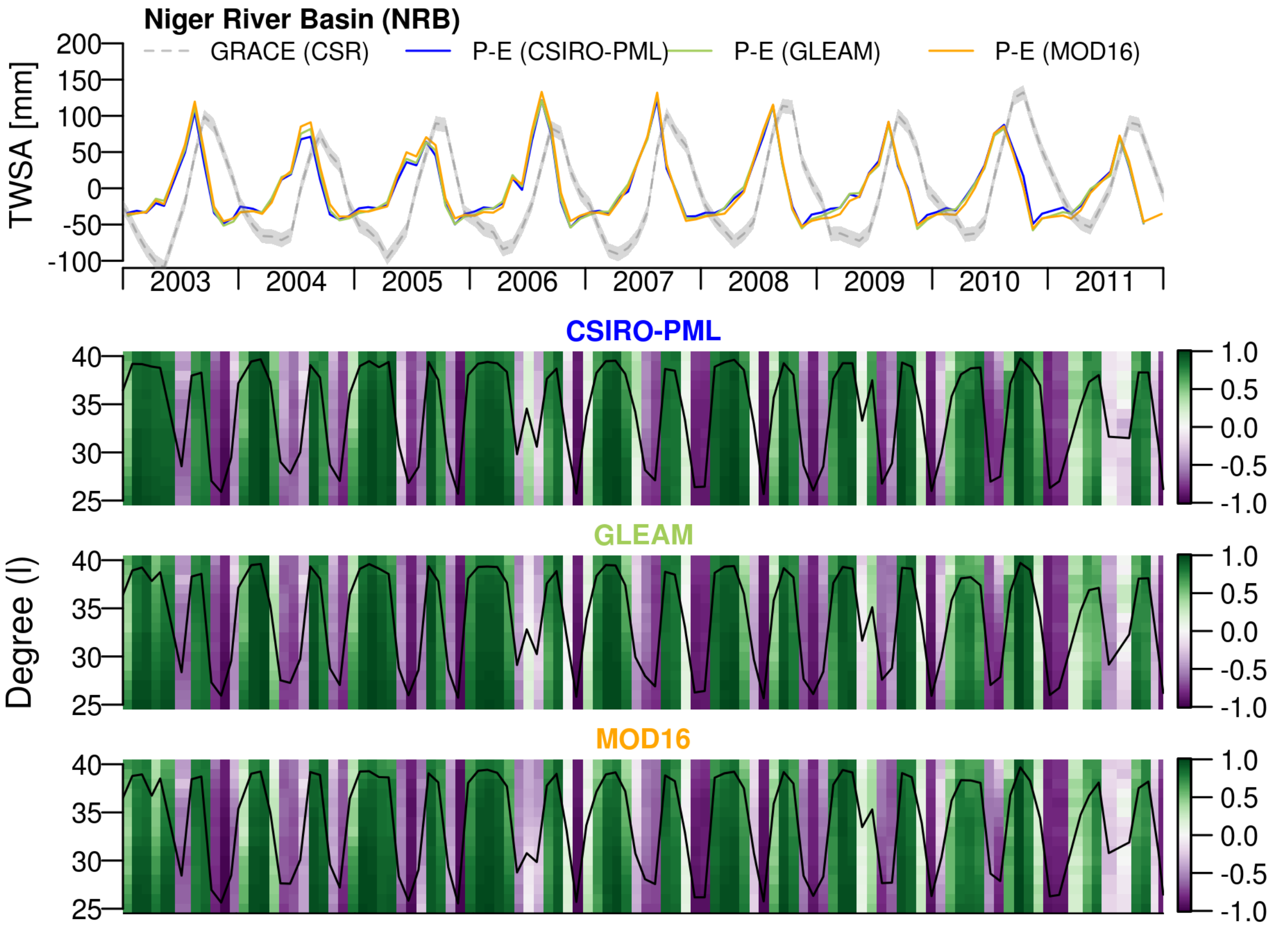

Figure 5. Top: anomalies of the terrestrial water storage (TWSA) observed by GRACE (with 20 mm uncertainty bounds) and P-E using three global evaporation products over the Niger river basin. Below: varying degree correlation measure (r_l) with time and degree (from 25 to 40) using the three global evaporation 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.

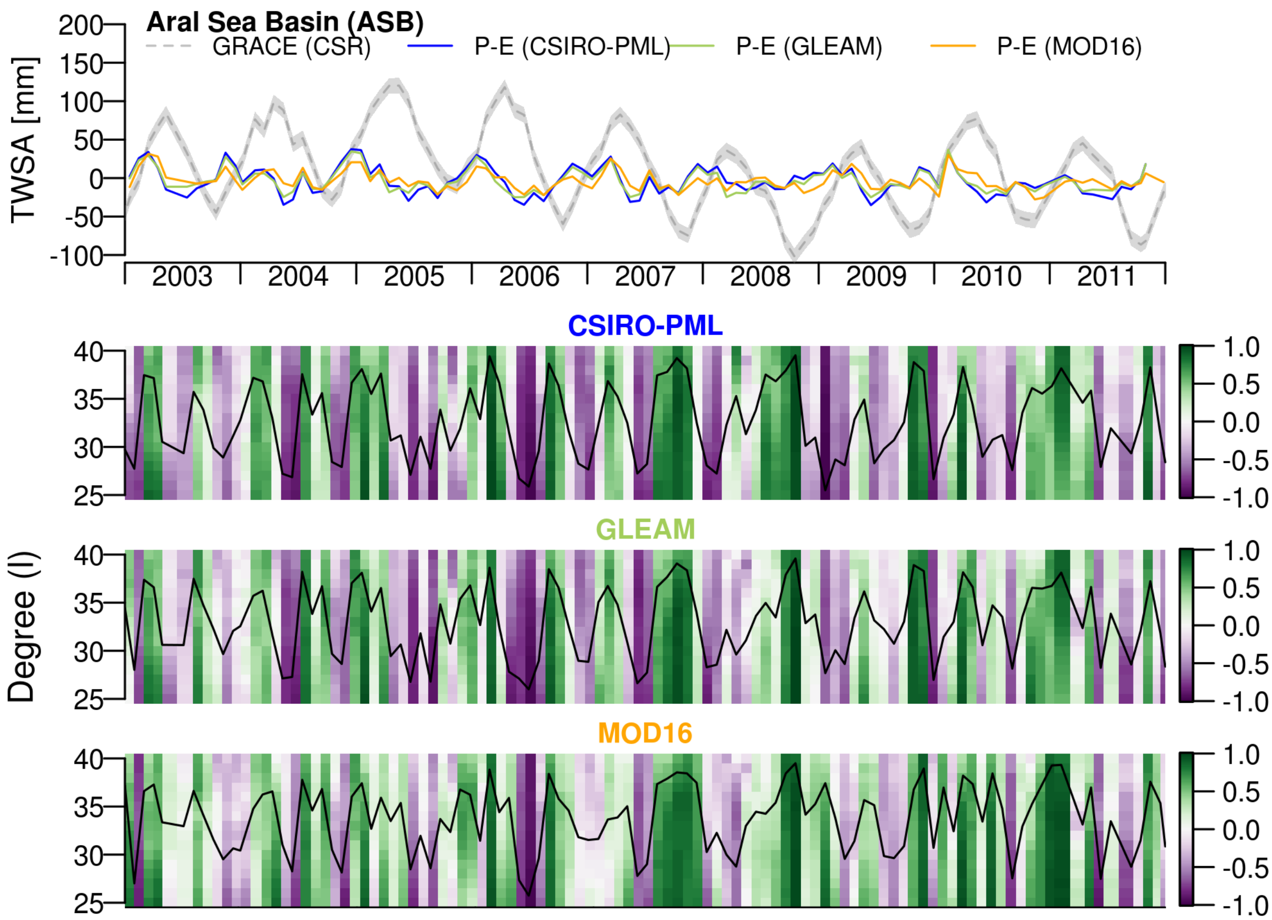

Figure 6. Top: anomalies of the terrestrial water storage (TWSA) observed by GRACE (with 20 mm uncertainty bounds) and P-E using three global evaporation products over the Aral basin. Below: varying degree correlation measure (r_l) with time and degree (from 25 to 40) using the three global evaporation products. The average r_l is shown as a time series (black line). The degree correlation measure can range from -1 to 1 as shown in the color scale on the secondary axis.

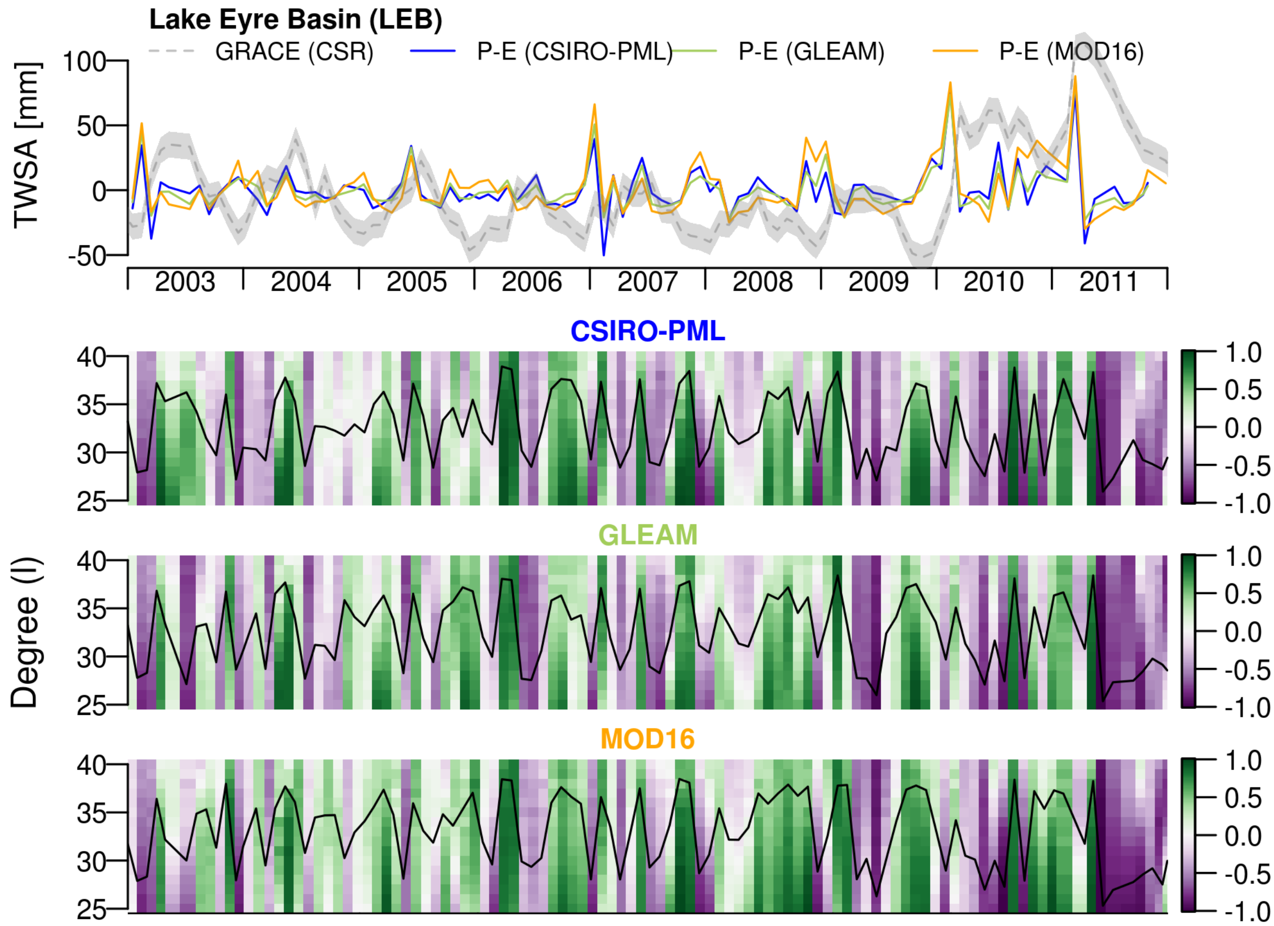

Figure 7. Top: anomalies of the terrestrial water storage (TWSA) observed by GRACE (with 20 mm uncertainty bounds) and P-E using three global evaporation products over the Lake Eyre basin. Below: varying degree correlation measure (r_l) with time and degree (from 25 to 40) using the three global evaporation 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.

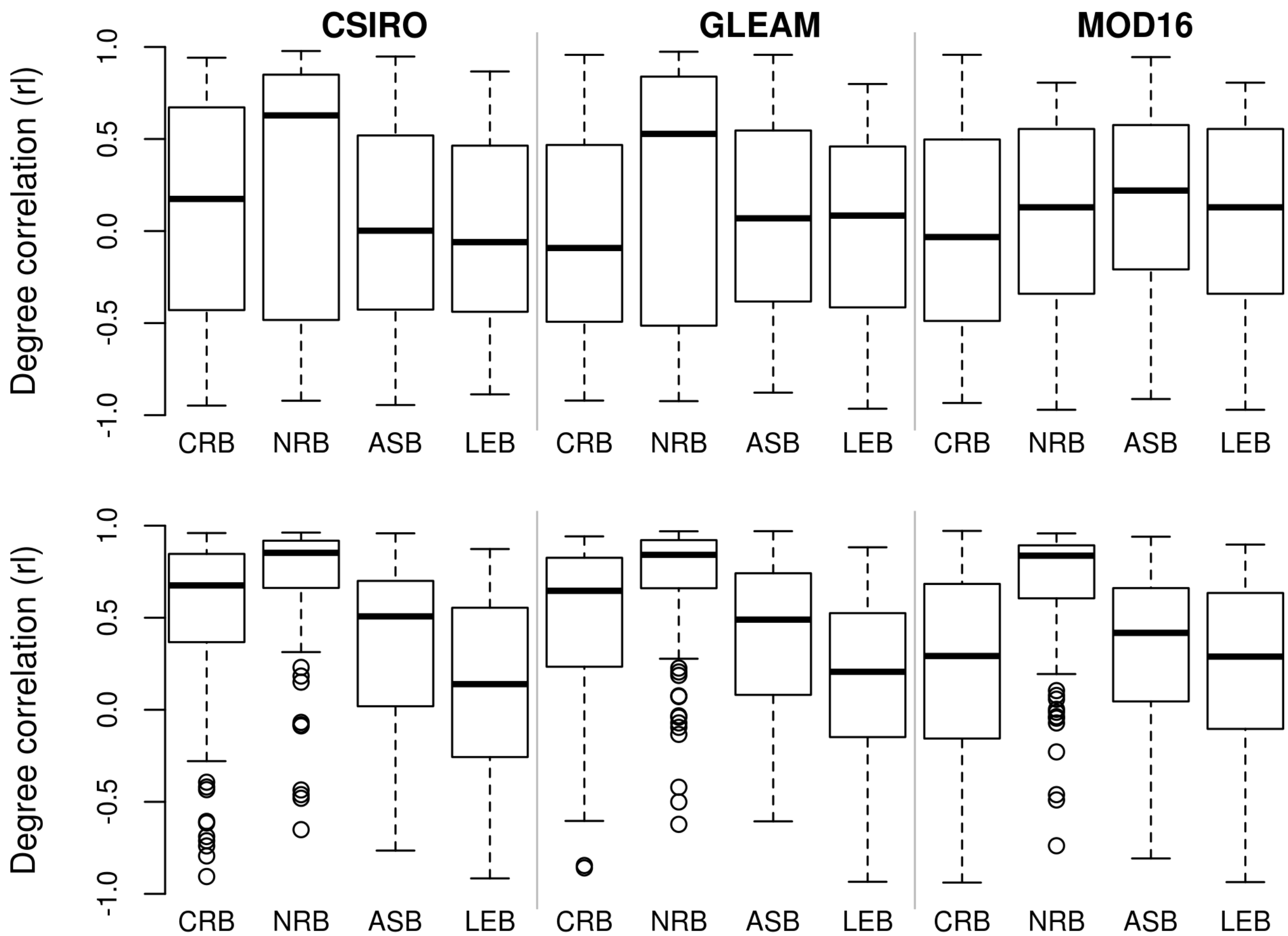

Figure 8. Top: average degree correlation statistics per study region and evaporation product. Bottom: GRACE data were shifted by two months to match the phase with P-E anomalies. The boxplots show the first, second (median) and third quartiles. Outliers, defined as data outside the 1.5 inter-quartile range (IQR) whiskers below or above the first and third quartiles are shown as circles.