Dear Professor Su,

Thank you for handling our paper.

We implemented the technical correction on the SEBS product provided by referee #1 as well as a short paragraph presenting results for the anomalies relative to the mean suggested by referee #3 (paragraph 3.4 and Appendix B).

We also updated Fig.3 and Fig.4 to update the color scale and clarify that some values go beyond the lower end of the colorscale.

We include the version of the manuscript with track changes below.

Kind regards,

Claire Michailovsky on behalf of the authors.

Investigating sources of variability in closing the terrestrial water balance with remote sensing

Claire I. Michailovsky[1], Bert Coerver[1,2], Marloes Mul[1], Graham Jewitt[1,3,4]

[1]IHE Delft Institute for Water Education, Delft, Netherlands
[2]now at Food and Agriculture Organization of the United Nations, Rome, Italy
[3]Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, Netherlands
[4]Centre for Water Resources Research, University of KwaZulu-Natal, Pietermaritzburg, South Africa

Correspondence to: Claire I. Michailovsky (c.michailovsky@un-ihe.org)

Abstract. Remote sensing (RS) data is becoming an increasingly important source of information for water resources management as it provides spatially distributed data on water availability and use. However, in order to guide appropriate use of the data, it is important to understand the impact of the uncertainties of RS data on water resources studies. Previous studies have shown that the degree of closure of the water balance from remote sensing data is highly variable across basins and that different RS products vary in their levels of accuracy depending on climatological and geographical conditions.

In this paper we analyzed the water balance derived runoff from global RS products for 937 catchments across the globe. We compared time-series of runoff estimated through a simplified water balance equation using 3 precipitation (CHIRPS, GPM and TRMM), 5 evapotranspiration (MODIS, SSEBop, GLEAM, CMRSET and SEBS) and 3 water storage change (GRACE-CSR, GRACE-JPL and GRACE-GFZ) RS datasets with monthly in situ discharge data for the period 2003-2016. Results were analyzed through the lens of 10 quantifiable catchment characteristics in order to investigate correlations between catchment characteristics and the quality of RS based water balance estimates of runoff, and whether specific products performed better than others in certain conditions.

The median Nash Sutcliffe Efficiency (NSE) for all gauges and all product combinations was -0.02, and only 44.9% of the time-series reached positive NSE. A positive NSE could be obtained for 73.7% of stations with at least one product combination, while the overall best performing product combination was positive for 58.4% of stations. This confirms previous findings that the best performing products cannot be globally established. When investigating the results by catchment characteristic, all combinations tended to show similar correlations between catchment characteristics and quality of estimated runoff, with the exception of combinations using MODIS ET for which the correlation was frequently reversed. The combinations with the GPM precipitation product performed generally worse than the CHIRPS and TRMM data. However, this can be attributed to the fact that the GPM data is available at higher latitudes compared to the other products, where performance is generally poorer. When removing high latitude stations, this difference was eliminated and GPM and TRMM showed similar performance.

The results show the highest positive correlation between highly seasonal rainfall and runoff NSE. On the other hand, increasing snow cover, altitude and latitude all decreased the ability of the RS products to close the water balance. The catchment's dominant climate zone was also found to be correlated with time series performance with the tropical areas providing the highest (median NSE=.11) and arid areas the lowest (median NSE=-0.09) NSE values. No correlation was found between catchment area and runoff NSE. The results highlight the importance of further studies on the uncertainties of the different data products and how these interact when combining them, as well as new approaches to using the data rather than simple water balance type approaches. Efforts to improve specific satellite products can also be better targeted using the results of this study.

1 Introduction

With increasing global population and pressure on the available water resources, it is increasingly important to understand the spatial and temporal distribution of water resources availability and use. Quantifying the components of the water balance is a necessary first step in sustainably managing resources in a river basin or catchment. However, the data available in many river basins is insufficient to make informed water management decisions. Global monitoring of discharge, which is one of the key variables of interest to water managers, has been in decline since the 1980s (Vorosmarty et al., 2001). In addition, even where in situ data exists, accessibility of the data can be problematic.

This data gap is increasingly being filled by remote sensing products which provides many advantages (see e.g. Sheffield et al., 2018 for a full review). For instance, remote sensing data can give valuable insights into the spatial variability of water availability and consumption which can be difficult or impossible to obtain through in situ data collection. Utilizing the hydrological variables currently derived from remote sensing, it is now theoretically possible to close the water balance and estimate runoff at the regional to global scale. However, due to uncertainties and errors in remote sensing data, this cannot currently be achieved at the scales and precision necessary for decision making (Sheffield et al., 2018).

Runoff estimation using remote sensing is typically done using some form of the following water balance equation (Eq.1) (see e.g. Syed et al., 2005):

$$R_o = P - ET_a - \frac{dS}{dt} \qquad (1)$$

where Ro is total runoff, P is the precipitation, ETa is the actual evapotranspiration and dS/dt is the total water storage change. Of the quantities in equation (1), all but the total runoff, which includes surface and subsurface components, can be derived from remote sensing at the global scale: remote sensing precipitation has been available for many years and is routinely used as input to hydrological models (see e.g. Stisen and Sandholt, 2010), ETa is not a direct RS measurement but many different algorithms have been developed to produce global scale ETa from RS data (Zhang et al., 2016), and total water storage change can be monitored using measurements of the variation of the Earth's gravitational field by the Gravity Recovery and Climate Experiment (GRACE, Wahr et al., 2004). We note that given adequate auxiliary information (such as for example bathymetry or rating curves), discharge can be monitored using radar altimetry (see e.g. Kouraev et al., 2004; Michailovsky et al., 2012). However, currently (2023) neither the radar altimetry nor the auxiliary information is available consistently at the global scale and in situ or modeled data is therefore necessary in order to assess the closure of the water balance using Eq.1.

A common approximation made when analyzing the terrestrial water budget using remote sensing over a hydrological basin or sub-catchment is to equate the total runoff with the discharge leaving the area of study. This is equivalent to the assumption that subsurface fluxes in and out of the basin are negligible. While this assumption is likely to have an impact, in particular for studies at small spatial scales (see e.g. Bouaziz et al., 2018; Fan and Schaller, 2009), it allows for the use of in situ discharge data to evaluate reliability of the remote sensing inputs to Eq. 1 which is then rewritten as Eq. (2):

$$Q = P - ET_a - \frac{dS}{dt} \qquad (2)$$

For the components of the water cycle which are available through RS, various datasets are available and each product is subject to uncertainties and errors. These include the fact that most remote sensing measurements are indirect, therefore requiring interpretation and calibration, subject to interference (e.g. by cloud cover and topography) and limited in their spatial and temporal resolution relative to the phenomena measured. Each product uses its own algorithms, gap filling procedures parameterization and validation methods to produce the variable of interest. Studies have shown that there is a large variability between the different products for a single variable (e.g. Sahoo et al., 2011).

Previous studies have analyzed the closure of the water balance with remote sensing and other global datasets from the regional to global scale. The first of such studies was performed by Syed et al. (2005) who used the land-atmosphere water balance to estimate discharge over the Amazon and Mississippi River Basins using data from the European Centre for Medium-Range Weather Forecasts (ECMWF) and GRACE data to measure water storage change. They found that the total basin outflow was well correlated with observed streamflow in spite of phase (in the Amazon) and amplitude (in the Mississippi) discrepancies. Sheffield et al. (2009) also analyzed the water budget closure for the Mississippi and found that the RS-estimated discharge was greatly overestimated. Sahoo et al. (2011) estimated the water budget from remote sensing and in situ discharge gauges over 10 global river basins and found errors in the runoff estimates of the order of 5 to 25% of the mean annual precipitation

values. Both Sheffield et al. (2009) and Sahoo et al. (2011) concluded that the largest contributor to the lack of closure of the water balance were errors and biases in the precipitation products used.

At the global scale, one of the most comprehensive studies of the closure of the water balance from global products (including remote sensing products, products derived from gauges and models) was carried out by Lorenz et al. (2014). They compared the ability of combinations of 5 precipitation products (4 derived from gauges and 1 including RS and gauge measurements), 6 ET products (including MOD16 and GLEAM from RS) and 2 storage change solutions from GRACE (GFZ and CSR) over 96 catchments spread around the world. No single product combination was found to consistently outperform the others across catchments but catchments with high seasonality tended to show better results.

More recently, Lehmann et al. (2022) performed a similar analysis on 189 river basins covering 90% of the global land surface and analyzed combinations of 11 precipitation and 14 ET datasets and 11 runoff datasets (including data from land surface models, gauge products and reanalysis datasets) and compared the computed storage change to GRACE data. They found that 95% of basins had a positive NSE for at least one product combination. They considered two catchment characteristics in analyzing their results and found that while no correlation between catchment area and closure of the water balance could be found, there was a correlation between climatic zone and performance for some of the datasets considered.

Other studies compared runoff computations obtained from different remote sensing input datasets to assess the best product combinations in specific regions. For example, Moreira et al. (2019) computed runoff using eq. 2 over South America using 2 precipitation products (TRMM and MSWEP), 2 ET products (MOD16 and GLEAM) and 3 storage change solutions from GRACE (CSR, JPL and GFZ) and found that using GLEAM for ET estimation and MSWEP for precipitation produced the best results. They also reported that greater biases were found in semi-arid basins with low runoff coefficients.

Following the findings from previous studies that different catchment characteristics (e.g. climate and seasonality) and different product combinations produced different results, this study aims to investigate both the ability of different combinations of RS products to reproduce in situ measurements of discharge, and to identify catchment characteristics that affect how well the closure of the water balance can be achieved among a wider range of catchment characteristics than those considered in previous studies. This is important in order to help water practitioners choose between different remote sensing datasets as the use of RS becomes more widespread in water balance assessments as well as to better understand the sources of uncertainties present in the different products and identify areas of improvement. In order to do this, 45 combinations of RS products (3 precipitation products, 5 ET products and 3 water storage change products) were used as input to the water balance equation (Eq. 2) and the discharge values computed were compared to discharge data collected from the Global Runoff Data Center (GRDC, 2019) over approximately 591 catchments (the number of catchments analyzed for each product combination varied due to coverage extent differences between products). The results were then analyzed using 10 quantifiable catchment characteristics to identify potential drivers of the goodness of fit between computed and in situ values.

2 Methodology

The ability of different remote sensing product combinations to correctly close the water balance was assessed by deriving runoff time-series for each combination of products using the water-balance equation of a river-basin (see Eq. 2) and comparing these RS-derived runoff values with monthly time step discharge measurements obtained from the Global Runoff Data Centre (GRDC) for a period of 14 years for which the RS products are consistently available.

The main drivers for the goodness of fit between calculated and observed runoff were investigated by evaluating 10 quantifiable basin characteristics.

2.1 Remote Sensing data

The data needed to solve the water balance for runoff are total water storage change, precipitation and actual evapotranspiration (see Eq. 2) over the study period. These time series were acquired from a variety of global remote-sensing products: three different precipitation products, five actual evapotranspiration products and three total water storage change products. An overview of these products is shown in Table 1 and details of the products are provided in the following sections.

Data was collected for a period of 14 years between 2003 and 2016, which are the full years for which the storage change from the Gravity Recovery and Climate Experiment (GRACE) data is available. All the products used are available within this timeframe, except for CMRSET, which was discontinued at the end of 2012.

The products cover most of the globe (see spatial coverage in Table 1). CHIRPS and TRMM do not cover areas north of 50° N and south of 50° S, meaning that Antarctica and the northern parts of Canada and Russia are excluded. The spatial extent of SSEBOP is also limited to areas between 80° N and 60° S. Furthermore, it is important to note that SEBS has many missing pixels, mainly over the larger deserts, such as the Sahara and the Arabian Desert, as well as the Taiga in Canada and Russia.

All the products were re-sampled to a monthly time-scale and to a spatial resolution of 0.05° (specific methods are detailed in in the following sections) and pixel values were weighted by area before computing the time-series to account for the changing pixel areas at different distances from the equator. The analysis focused on spatial aggregates of runoff for catchments larger than 10,000 km^2 and the spatial resampling was therefore not expected to have a large impact on the results. For studies which focus on smaller scales or at the pixel-level, the impact of spatial resampling would need to be carefully considered. The choice of a monthly time scale was motivated by the timescales of the available remote sensing, in particular the GRACE dataset.

2.1.1 Precipitation

Different sensors and algorithms are used to estimate global precipitation from remote sensing. Many of the available precipitation products combine measurements from sensors aboard multiple satellites in order to be able to achieve higher temporal resolutions and some products are merged with in situ gauge data to improve accuracy (Sheffield et al., 2018). In this study, the following three products were used:

- The Tropical Rainfall Measuring Mission (TRMM) Multi-satellite Precipitation Analysis (TMPA) 3B42 product (Huffman et al., 2007).
- The Climate Hazards group Infrared Precipitation with Stations (CHIRPS) version 2 product (Funk et al., 2015).
- The Global Precipitation Measurement (GPM) mission Integrated Multi-satellitE Retrievals for GPM (IMERG) Final Run (Huffman et al., 2019).

The datasets had to be resampled from their native resolutions (see Table 1) to obtain monthly data at 0.05° spatial resolution:

- The TRMM TMPA and GPM IMERG products were resampled to 0.05°using the nearest neighbor method.
- The daily TRMM and CHIRPS daily data products were summed to obtain monthly values.

It should be noted that the products used are in large part computed from the same source satellite measurements. In particular, while the core GPM satellite was launched in February 2014, the IMERG algorithm was used to extend the time series back to June 2000 using data from the TRMM satellite to produce a continuous long-term dataset. The TRMM satellite stopped operating in 2015 and, post 2015, the TMPA algorithm was applied to GPM data in order to continue producing data (Huffman, 2020).

170

Table 1: Overview of the different remote-sensing products acquired

Product (version)	Availability	Spatial Resolution	Temporal Resolution	Spatial Coverage	Reference	Obtained from:
Precipitation						
CHIRPS (v2)	1981-present	0.050°	Daily	50° S-50° N	Funk et al. (2015)	https://data.chc.ucsb.edu/products/CHIRPS-2.0/
TRMM TMPA (3b42 v7)	1998-2020	0.25°	Daily	50° S-50° N	Huffman et al. (2007)	https://disc2.gesdisc.eosdis.nasa.gov/opendap/TRMM_L3/TRMM_3B42_Daily.7/
GPM 3IMERGDF (v06)	2000*-present	0.10°	Monthly	90° N-90° S	Huffman et al. (2019)	https://gpm1.gesdisc.eosdis.nasa.gov/opendap/GPM_L3/GPM_3IMERGDF.06/
Evapotranspiration						
MOD16 A2 (v006)	2001-present	500m	8-Daily	90° N-90° S	Mu et al. (2011)	Google Earth Engine image collection: MODIS/006/MOD16A2
SSEBOP (v4)	2003-present	0.010°	Dekadal	80° N-60° S	Senay et al. (2013)	https://edcintl.cr.usgs.gov/downloads/sciweb1/shared/fews/web/global/monthly/eta/downloads/
GLEAM (v3.3b)	2003-2018	0.25°	Daily	90° N-90° S	Miralles et al. (2011)	sftp://hydras.ugent.be (access instructions: https://www.gleam.eu/ - current version: v3.6b)
CMRSET	2003-2012	0.050°	Monthly	90° N-90° S	Guerschman et al. (2009)	Shared by Dr. Guerschman
SEBS (5km Global ~~Monthly~~ Daily ET)	2000-2017	0.050°	Monthly	90° N-90° S	(Chen et al., 2021)	Obtained from: https://data.tpdc.ac.cn/en/data/df4005fb-9449-4760-8e8a-09727df9fe36/
Water storage change						
GRACE CSR (TELLUS_GRAC_L3_CSR_RL06_LND v6.0)	2003-2017**	1.0°	Monthly	90° N-90° S	Landerer (2019a)	Retired product – see: https://podaac.jpl.nasa.gov/dataset/TELLUS_GRAC_L3_CSR_RL06_LND
GRACE GFZ (TELLUS_GRAC_L3_GFZ_RL06_LND v6.0)	2003-2017**	1.0°	Monthly	90° N-90° S	Landerer (2019b)	Retired product – see: https://podaac.jpl.nasa.gov/dataset/TELLUS_GRAC_L3_GFZ_RL06_LND
GRACE JPL (TELLUS_GRAC_L3_JPL_RL06_LND v6.0)	2003-2017**	1.0°	Monthly	90° N-90° S	Landerer (2019c)	Retired product – see: https://podaac.jpl.nasa.gov/dataset/TELLUS_GRAC_L3_JPL_RL06_LND

*The TRMM mission ended in 2015, but the TMPA product continued to be produced using data from GPM, the GPM satellite was launched in 2015 but the IMERG product starts in 2000, using TRMM data. **The GRACE mission produced data until July 2017, the GRACE-FO satellite started producing data from June 2018.

2.1.2 Evapotranspiration

Evapotranspiration (ET) obtained from RS data is not a direct measurement, and many different inputs are required for models to be able to represent the biophysical and environmental controls on ET (see e.g. Zhang et al., 2016). Five different evapotranspiration products have been used to solve the water balance for runoff in this study[1].

- The Operational Simplified Surface Energy Balance (SSEBop, Senay et al., 2013).
- CSIRO MODIS Reflectance-based Evapotranspiration (CMRSET, Guerschman et al., 2009).
- Global Land Evaporation Amsterdam Model (GLEAM, Miralles et al., 2011).
- Surface Energy Balance System (SEBS, Chen et al., 2021).
- MODIS Global Terrestrial Evapotranspiration Algorithm (MOD16, Mu et al., 2011).

These products use different methods and data sources for estimating evapotranspiration rates. For example, the MOD16 algorithm is based on the Penman-Monteith equation, CMRSET and GLEAM use modified versions of the Priestly–Taylor equation while SSEBop and SEBS use surface energy balance approaches. More detail can be found in the publications listed for each product.

In order to obtain monthly data at 0.05° spatial resolution from the resolutions listed in Table 1 the following was done:

- The daily and dekadal fluxes from SSEBOP and GLEAM were summed to obtain monthly values.
- The 8-daily data from MOD16 were summed to monthly values (with reduced weights for images partially within a specific month). Missing data within a month was filled by setting the missing data to the monthly average of the available 8-day evapotranspiration in that month.
- MOD16, SSEBop and GLEAM were resampled to 0.05° using the nearest neighbor method.

2.1.3 Storage Change

Total water storage (the sum of surface and subsurface water storage) cannot be directly measured from remote sensing. However, Total Water Storage Anomalies (TWSA), i.e. the deviation in total water storage relative to the long term mean, can be obtained from the Gravity Recovery And Climate Experiment (GRACE) satellites which maps the Earth's gravity field approximately every 30 days (Biancamaria et al., 2019).

The TELLUS GRACE Level-3 Monthly LAND Water-Equivalent-Thickness Surface-Mass Anomaly Release 6.0 products from three processing centers were used in this study (Landerer and Swenson, 2012):

- the University of Texas – Center for Space Research (CSR, Landerer, 2019a)
- Geo Forschungs Zentrum (GFZ, Landerer, 2019b)
- Jet Propulsion Laboratory (JPL, Landerer, 2019c)

[1] Two other products were considered before being excluded from the study: the WaPOR dataset as it does not yet have global coverage, and ALEXI as it was not available to the authors at the time of the study.

GRACE data is available between January 2003 and July 2017. The data is available in quasi-monthly time steps with variable windows of observation. However, most of the data is centered on the 16th of each month. The data was interpolated to the 16th day of every month and the central difference method was used to calculate the change in storage (see e.g. Biancamaria et al., 2019). Finally, the data was resampled to 0.05° using the nearest neighbor method.

2.2 In situ data: Global Runoff Data Centre

The RS-derived runoff was validated using observed runoff from the Global Runoff Data Centre (GRDC), whose dataset comprises more than 9,900 gauging stations all over the world. By filtering to identify stations with an upstream catchment larger than 10,000 km^2 and at least one record after January 1st 2003, an initial selection of 1,149 gauging stations was made. A large number of these stations are located in northern America, while the rest are spread out across the other continents (see Figure 1). Unfortunately, among the selected stations, there are very few stations located in some parts of the world, in particular Northern Africa, Central Asia and Southern Asia.

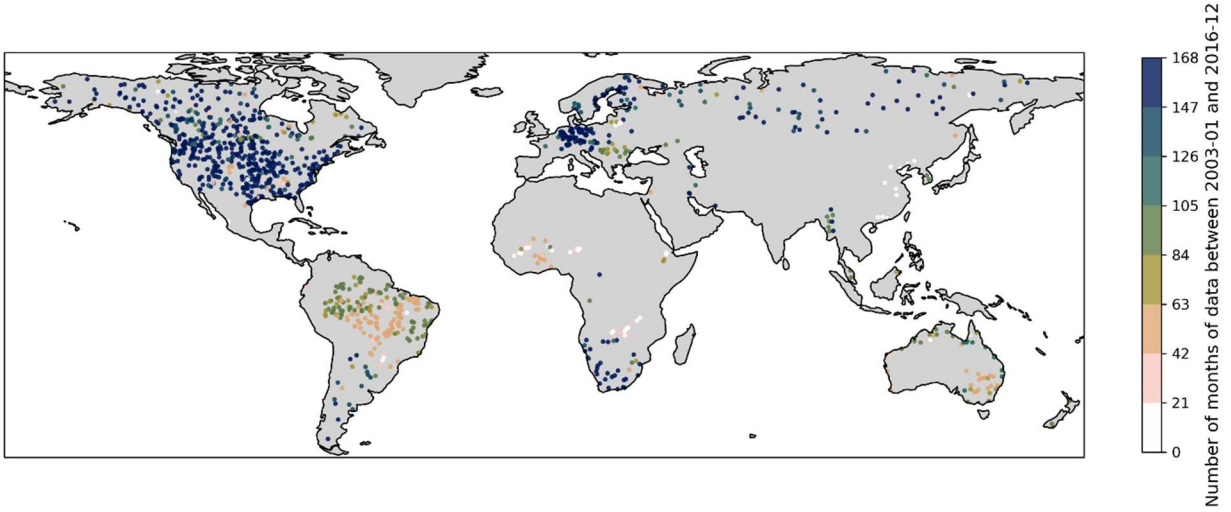

Figure 1: Locations of the acquired GRDC stations with runoff data.

Within the period 2003-2016, the selected stations have an average of 125 months of data, with just over half (515 stations) having more than 160 months of data out of a maximum possible of 168 months. For the first five years of this period nearly all the selected stations have data with an average of 1015 data points available each month. After 2008, the availability starts to decrease and by 2008, the average number of data points per month drops to 580. A total of 143,117 monthly runoff records were used for the analysis.

Watershed boundaries were also obtained from the GRDC (GRDC, 2011). The largest catchment covers 4,680,000 km^2 (the Amazon River), and most of the catchments (862) are between 10,000 and 93,600 km^2. The mean catchment size is 141,259 km^2. Altitude was known for 764 of the stations, and mean station altitude is 298.4 m.a.s.l. with a large number (161) of stations being located at altitudes below 40 m.a.s.l.

Many river basins contain multiple GRDC stations, meaning that among the 1,149 selected stations some represent nested catchments.

The monthly mean GRDC data is given in m^3/s, and was converted to mm/month in order to be compared to the monthly runoff computed from remote sensing data. This was done by dividing by the catchment area.

2.3 Runoff time-series from remote sensing

Solving the water-balance for the different combinations of three precipitation, five actual evapotranspiration and three storage-change products, results in a total of 45 solutions. Each of these solutions consists of a series of maps of the RS-derived runoff in mm/month. For each GRDC station, the RS derived runoff time series is obtained by averaging the pixels within the corresponding catchment.

Extracting these time-series at the 1,149 locations of the selected GRDC stations from these 45 combinations gives, 51705 time-series to analyze.

In practice the number of time series analyzed was lower due to several issues. First of all, calculated time-series that have fewer than 30 matching data points with the GRDC data were omitted. Secondly, some of the selected stations (or their catchments) are (partially) located outside of the coverage area of some of the products (see Table 1). Finally, months for which more than 20% of the pixels in a catchment were missing have been excluded (no gap-filling has been done), occasionally leading to the loss of an entire times-series (for example, as mentioned previously, SEBS has many missing pixels in some parts of the world). This finally resulted in 937 locations with sufficient data and 31734 time series.

2.4 Validation

The computed monthly runoff time-series have been compared with the GRDC data through the Nash–Sutcliffe model efficiency coefficient (NSE). The NSE is defined as (Nash and Sutcliffe, 1970):

$$NSE = 1 - \frac{\sum_{t=1}^{T}(Ro_c^t - R\ \ _0^t)^2}{\sum_{t=1}^{T}(Ro_0^t - \overline{Ro_o})^2} \qquad (3)$$

where $\overline{Ro_0}$ is the mean of the observed runoffs, Ro_c^t is the RS-derived runoff at time t and Ro_0^t is the observed runoff at time t.

2.5 Catchment Characteristics

We selected 10 RS derived catchment characteristics based on the findings of earlier studies to investigate correlations with quality of RS estimates of discharge. These are summarized in Table 2 and detailed below.

Table 2: Catchment characteristics considered in this study

Description (continuous/discrete)	Abbreviation	Unit	Data Source

Size of the catchment (continuous)	Area	km^2	GRDC (GRDC, 2019)
Distance of the catchment outlet to the equator (continuous)	|Latitude|	DD	GRDC (GRDC, 2019)
Altitude of the catchment outlet (continuous)	Altitude	m.a.s.l	GMTED10
Total dam storage capacity in the catchment (continuous)	S_{dam}	$10^6 m^3$	GRAND (Lehner et al., 2011)
Seasonality: Standard deviation of the monthly precipitation in the catchment (continuous)	SDP	mm/month	GPM (Huffman et al., 2019)
Ratio between the mean annual runoff and the total dam storage capacity (continuous)	$\overline{Ro_{yearly}}$: S_{dam}	−	GRAND, GRDC
Mean ratio between the monthly runoff and precipitation (continuous)	R_o : P	−	GRDC, GPM
Mean of the temporal and spatial snow-cover (continuous)	$\overline{NDSI}$	%	MOD10 (Hall et al., 2006)
Dominant land cover class (discrete)	LC	-	GlobCover2009 (ESA and UCLouvain, 2010)
Dominant climate class (discrete)	Climate	-	Köppen-Geiger Classification (Beck et al., 2018)

Catchment area was chosen as a catchment parameter as it is expected that in larger catchments, the random errors may be compensated by averaging over large areas. Beyond this, the resolution of the GRACE product should also allow for better performance over larger catchments. While Biancamaria et al. (2019) found that GRACE could provide good estimates of storage change for catchments larger than 50,000 km^2, most studies have considered only very large basins (>100,000 km^2). *Latitude* of the outlet of the catchment (or the distance to the equator in degrees) and *snow cover* were both chosen because precipitation products are known to have higher uncertainties at high latitudes and in the estimation of snow than in that of liquid precipitation (Tian and Peters-Lidard, 2010). Snow storage also adds a storage and therefore lag to the runoff generated in the basin which, while it should be captured by the GRACE data, can add another layer of uncertainty. ET products, in particular those based on measurements of land surface temperature, may also face issues in computing sublimation (Xu et al., 2019).

The *altitude* of the catchment outlet is evaluated to see any difference in accuracy between river catchments with an outlet at sea level and sub-catchments with an outlet at a higher altitude. Altitude of catchment outlet is also used as a proxy for topography and precipitation products are known to have higher uncertainty over areas of rough topography (Tian and Peters-Lidard, 2010).

Dam storage capacity was also considered due to the smoothing effect on the runoff. While the dam storage should be captured by the GRACE data, it has been shown that GRACE solutions do not always correctly locate the relatively punctual changes

in storage due to signal leakage which could impact the results (Wang et al., 2019). Dam storage capacity relative to mean annual runoff was also considered both as a measure of the level of modification of the basin, and as normalization for total dam storage capacity.

The *seasonality of rainfall* varies greatly around the world. Some regions have a clear dry and wet season, while others receive rainfall throughout the entire year. In order to make a distinction between these different rainfall patterns, the standard deviation of the monthly rainfall was chosen as a parameter. A catchment with a clear wet and dry season will have a higher standard deviation than a catchment with precipitation throughout the year.

Finally, the *ratio between runoff and precipitation* is considered. Catchments with a low runoff to precipitation ratio will typically have a high evapotranspiration rate relative to precipitation, while a higher ratio indicates a low evapotranspiration rate. Catchments with ratios above 1 indicate discharge originating from either storage depletion in the basin, or inter-basin transfers.

Besides the above characteristics which can be described by continuous variables, the following two discrete characteristics were considered:

The *dominant climate class* according to the Köppen-Geiger climate classification was computed for each catchment based on data from Beck et al. (2018). This was considered as previous water balance closure studies have shown variable performance under different climate conditions (e.g. Lorenz et al., 2014),

The final catchment characteristic considered was *dominant land cover class* in the catchment (computed from GlobCover2009 (ESA and UCLouvain, 2010)). This was considered due to the variable performance of ET products in over different land cover types (e.g. Senay et al., 2013).

For each of the continuous catchment characteristics, the Spearman Rank correlation coefficient, which is the Pearson correlation coefficient between the ranks of the variables, was computed to assess the correlation between each catchment characteristic and the NSE values of the discharge time series. The significance of the correlations ($p < 0.05$) was tested using a two-sided student t-test.

For the two non-continuous characteristics (LULC and Climate class), the influence of the characteristic on the performance was analyzed by comparing the NSE values obtained per class.

3 Results and Discussion

3.1 Results per GRDC station

NSE values were computed for the 45 possible product combinations, for all GRDC stations possible for each combination. Figure 2~~FIG~~ shows a histogram of the NSE values for all 31734 time-series computed and Figure 3 shows the median NSE value for all possible product combinations at each of the 937 GRDC stations for which at least one NSE value could be computed.

For all combinations of products at all available GRDC stations, 44.9% of the generated discharge time series achieve a positive NSE value, with only 3.4% ~~obtained a median~~obtaining an NSE > 0.5. When split by GRDC station, 36.9% of the stations achieve a positive median NSE value and 2.5% a median NSE of > 0.5. A positive NSE indicates a model performs better than the long-term mean of the observed time series as a predictor. Hydrological models are often considered to be of good quality when reaching NSE values of > 0.5, although many studies use different thresholds (Moriasi et al., 2007).

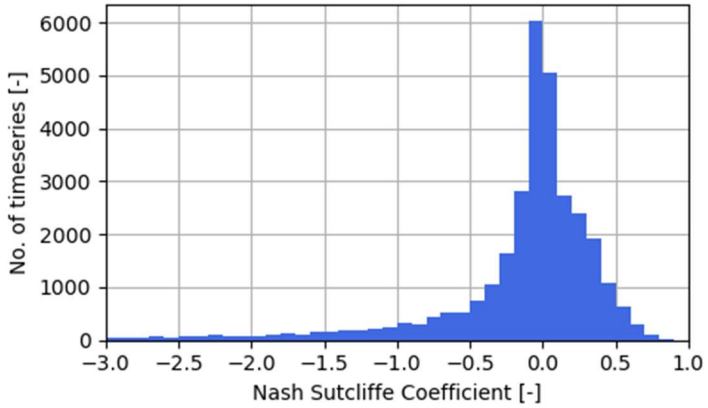

Figure 2: Distribution of NSE values for all time-series. 891 time series with NSE<-3 not shown (2.8% of timeseries)

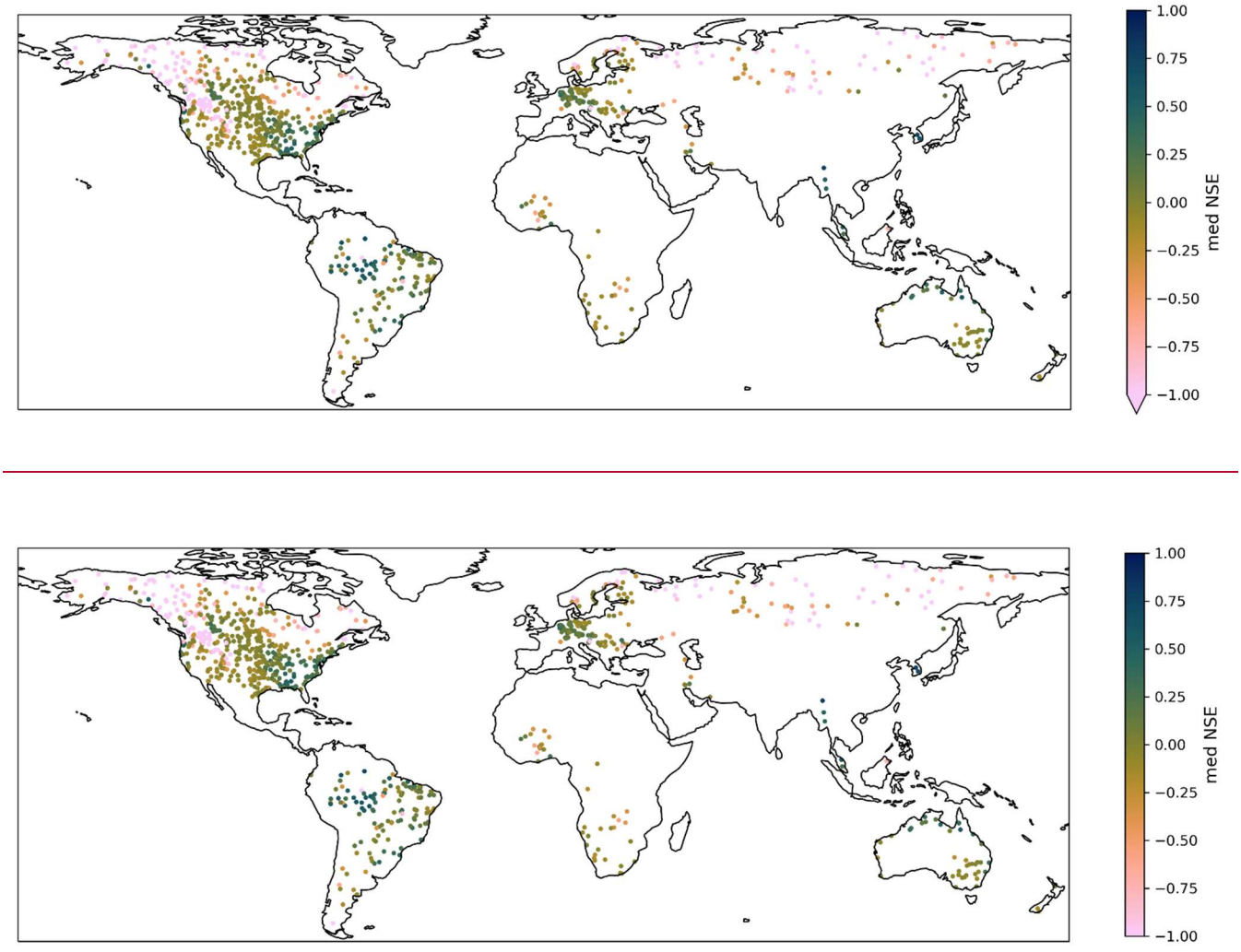

Figure 3: Median NSE for different product combinations at each GRDC station. 125 stations have a median NSE below -1, the color scale was cropped to -1 for legibility.

When considering the maximum NSE reached at each station, it was determined that a positive NSE was reached for at least one product combination for 73.7% of the stations, and an NSE of more than 0.5 was reached for 7.3% of the stations. The geographical distribution of maximum NSE values is shown in Figure 4.

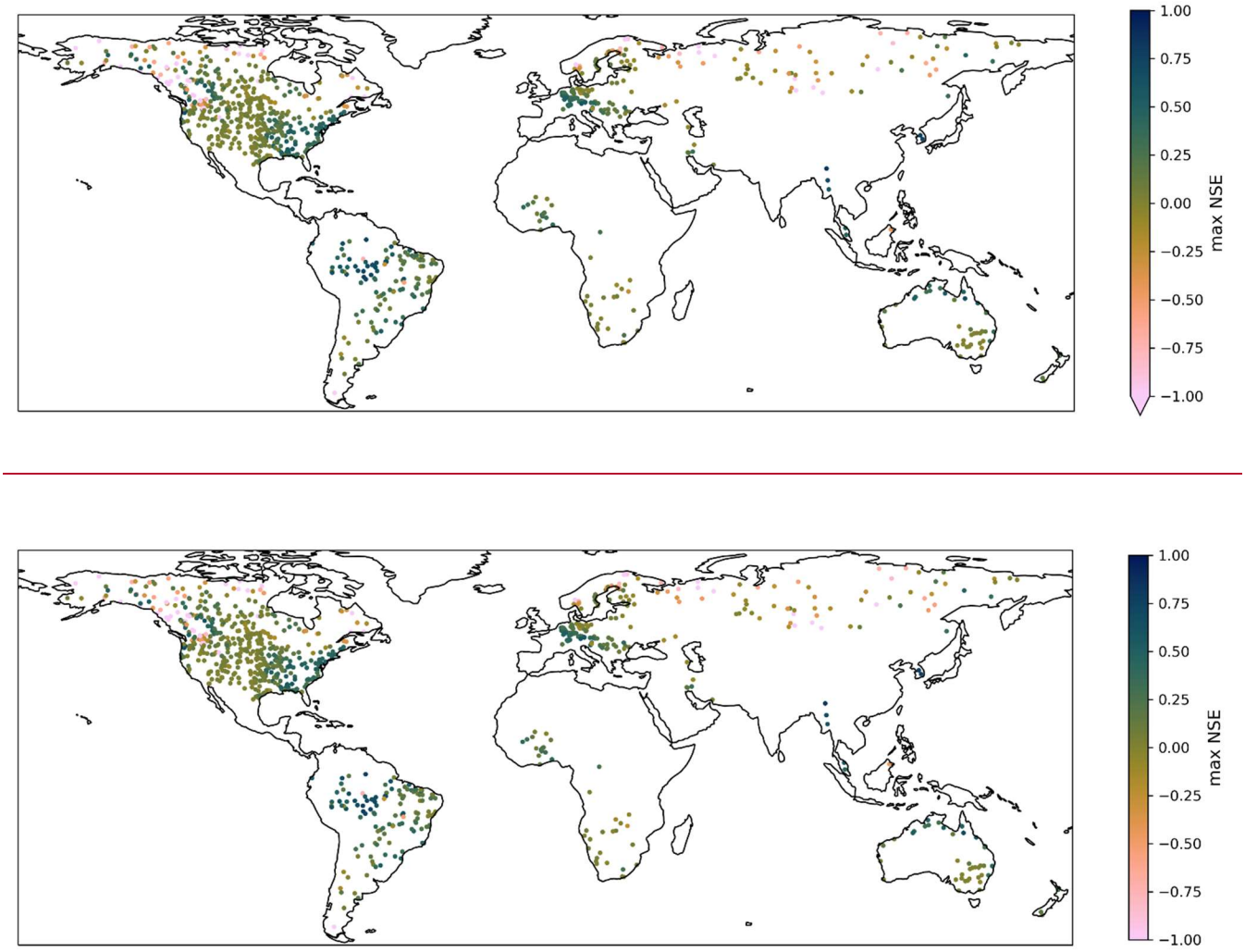

Figure 4: Max NSE achieved at each GRDC station. 43 stations have a maximum NSE below -1, the color scale was cropped to -1 for legibility.

In the studies performed by Lorenz et al. (2014), positive NSE values were reached in 29 of the 96 (30%) basin considered while in the study by Lehmann et al. (2022) this was achieved in 180 of 189 (95%) of the basins. These results are however difficult to compare directly due to the different products chosen and the different basins considered. In terms of the datasets considered, we chose to limit our study to remote sensing products, excluding land surface models, station based gridded products as well as reanalysis products. This differs from the two aforementioned studies as our goal is to specifically investigate the remote sensing products and work with independent datasets.

Our study, while it considers the largest number of catchments, was limited to those with GRDC station data available over our time period of interest which excluded some large basins. On the other hand, many smaller catchments were considered, including nested catchments where multiple stations were available. Areas with more dense gauging networks are therefore overrepresented in our study and these correlate with particular catchment characteristics (for instance climate zone) which can influence the ability of remote sensing to close the water balance as will be seen in Sect 3.3.

3.2 Results per product and product combination

For the product combinations based on the GPM rainfall product, an average of 925 time series NSE values could be calculated per combination, while for the combinations based on the TRMM and CHIRPS products, an average of 599 NSE values per combination could be calculated (due to the smaller spatial coverage of these TRMM and CHIRPS).

The median NSE values for all GRDC stations available for the 45 possible product combinations are presented in the Appendix ~~1~~A. The best performing combination was CHIRPS – SEBS – JPL which yielded 58% of positive NSE values while GPM – GLEAM – CSR/GFZ/JPL, yielded 35% of positive NSE values. Only 3.4% of the discharge time series generated reached the threshold of 0.5, with the best combination (CHIRPS - CMRSET – GFZ) reaching this value for 5.9% of stations. The worst performing combination (GPM - GLEAM – GFZ) reached NSE>0.5 for only 1.3% of stations.

In order to make the product combinations more comparable, the same results are presented for 1) all possible time series (columns A in Appendix ~~1~~A) and 2) for only those stations for which all products could be used (columns B in Appendix ~~1~~A). The main consequence of this is that the high latitude stations which are only covered by GPM are removed from the analysis which narrows the performance gap between GPM and other precipitation products.

Table 3 shows that the NSE of the computed discharge is most sensitive to the choice of ET product. With median NSE values ranging from -0.02 to 0.01. The ET product with the highest median NSE and number of NSE series with values above 0 is MOD16. The product with the highest number of series producing NSE values above 0.5 is SEBS (followed closely by SSEBop and CMRSET). For precipitation, the impact of different products on the overall median NSE is negligible when not considering high latitude stations where only GPM is available. GPM produces the highest number of series with NSE values above 0, while CHIRPS produce the highest number of series with NSE values above .5. The computed NSE was not found to be sensitive to the choice of GRACE solution used.

Table 3: Median NSE for time series containing specific products as well as percentage of time series with positive NSE, NSE above 0.5 (n. NSE>0.5) and total number of time series using the product (n. series). Series have been limited to those covered by all product combinations (591 GRDC stations).

Variable	Product	Median NSE	%NSE>0	%NSE>0.5	n. Series
P	TRMM	-0.00	49	3.2	8850
	GPM	-0.00	50	3.9	8850
	CHIRPS	-0.00	49	4.7	8850
ET	SSEBOP	-0.00	48	4.9	5310

	MOD16	0.01	52	3.2	5310
	SEBS	0.01	54	4.9	5310
	GLEAM	-0.02	43	2.0	5310
	CMRSET	-0.01	49	4.8	5310
GRACE	JPL	-0.00	50	3.9	8850
	CSR	-0.00	49	4.0	8850
	GFZ	-0.00	49	4.0	8850

The precipitation and ET products used in the best performing combination for each station are shown in Figure 5 and Figure 6. Because of the low sensitivity of NSE to storage change solution, no map was generated for the different storage change products.

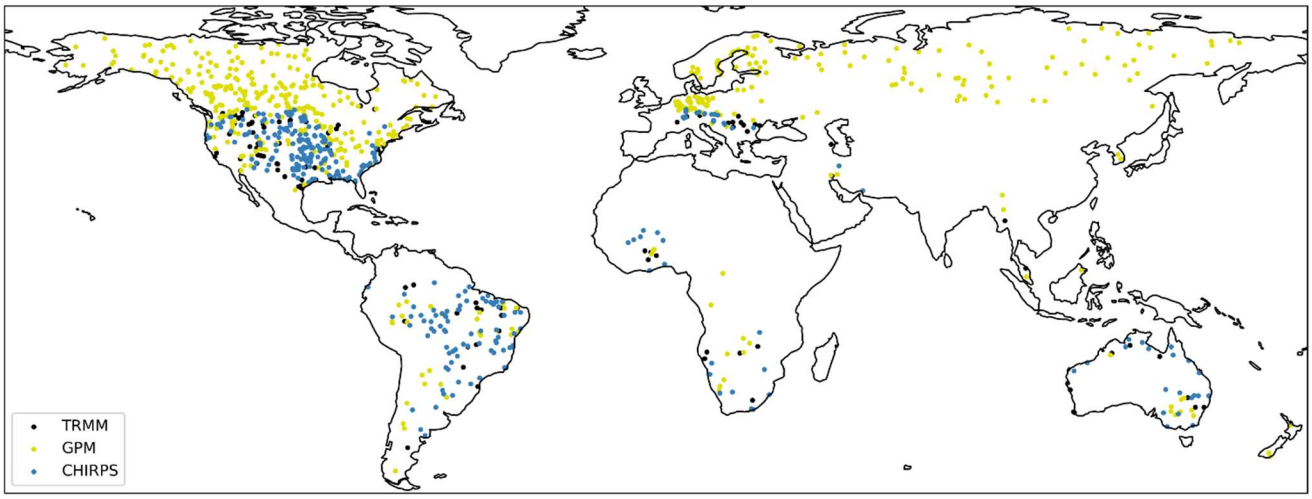

Figure 5: Precipitation product used in combination with highest NSE at station. Note that GPM is the only product available for latitudes >50°.

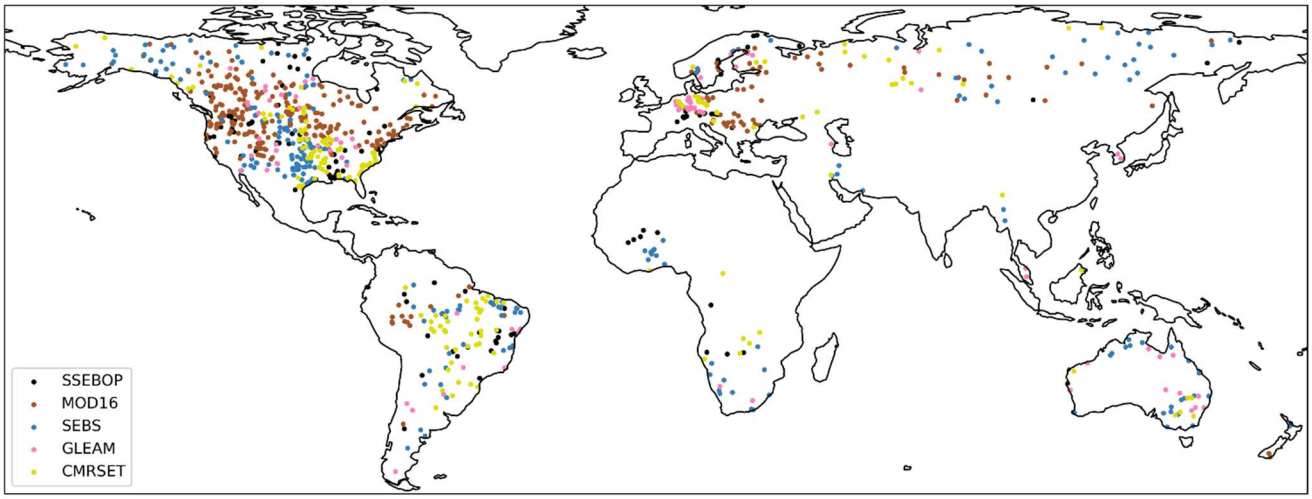

Figure 6: ET product used in combination with highest NSE at station

These results show that no single product or combination consistently outperformed others when it comes to the closure of the water balance. This is consistent with findings of previous studies (Lehmann et al., 2022; Lorenz et al., 2014). Some geographic patterns in the better performing products appear in Figure 5 and Figure 6 and will be discussed in the context of the catchment characteristics in the following section.

3.3 Results per catchment characteristic

For each of the continuous catchment characteristics listed in Table 2, correlations between the characteristic and the NSE at the GRDC station were computed. Figure 7 shows a summary of the correlations found for all product combinations and the catchment characteristics.

Presence or absence of correlation as well as whether the correlation strength and sign are consistent across most product combinations.

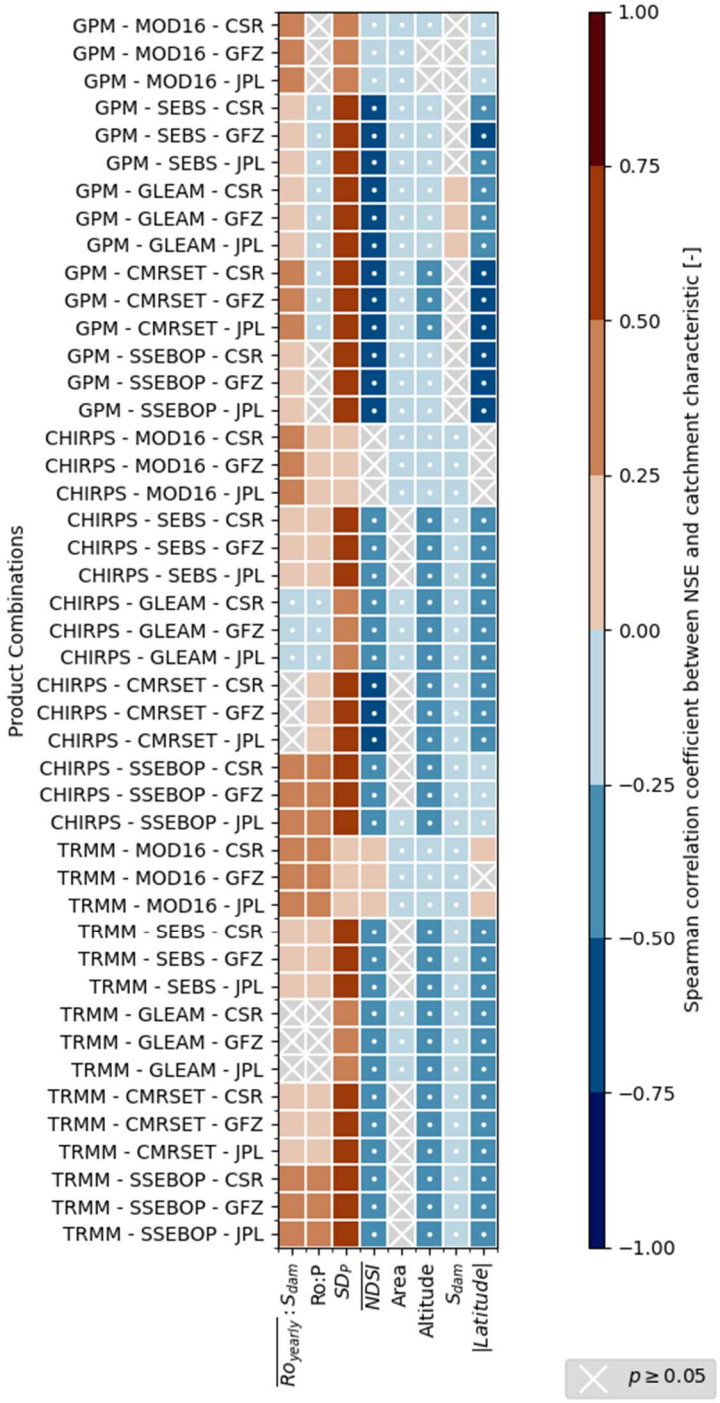

Figure 7: Spearman correlations for different product combinations between the NSEs of catchments and characteristics of those catchments. See Table 2 for an overview of the catchment characteristics. White dots were added to the negative correlations for monochromatic legibility.

Of the catchment characteristics described by a continuous variable, seasonality (SDp) shows the strongest correlation with the NSE of the discharge. All product combinations, showed a significant correlation with the standard deviation of precipitation. It should be noted that precipitation from GPM was used to compute seasonality, meaning that errors and uncertainties in GPM data could affect catchment classification. The influence of seasonality is in agreement with the findings of Lorenz et al. (2014) who found that the closure of the water balance can be better achieved in basins with a strong seasonal precipitation signal. Lorenz et al. (2014) observed that in catchments with low seasonal runoff variability, the biases in the different input datasets prevented the accurate computation of runoff.

Snow cover has the strongest negative correlation with NSE. NDSI shows a significant negative correlation for 39 of the 45 product combinations. Combinations including MODIS ET and CHIRPS or TRMM precipitation are the only ones for which no correlation or a positive correlation was found. Altitude at the gauging station, which is correlated to snow cover for smaller basins, shows a weaker negative correlation with NSE. The strong negative correlation with snow could be due to multiple factors. For instance, snow retrievals have lower accuracies as compared to liquid precipitation retrievals from satellite and precipitation retrievals are less accurate over frozen ground (Tang et al., 2020; Tian et al., 2014; Tian and Peters-Lidard, 2010), ET products may not capture the process of sublimation as well as other types of ET (see e.g. Xu et al., 2019), and the snow storage variations which drive discharge timing in some catchments may not be adequately captured by GRACE. Analysis of runoff versus discharge totals over hydrological years, rather than monthly could mitigate the snow storage issue. A similar analysis with more recent data should also be carried out to check if better results for catchments further from the equator (>50°N and >50°S) can be obtained, as the GPM data from the TRMM era (pre-2014) for higher latitudes is considered partial coverage. The GPM core observatory also has higher sensitivity to snowfall than earlier sensors (Behrangi et al., 2018) and was only launched in 2014.

Latitude also shows a correlation with NSE for 39 of 45 product combinations while the remaining 6 show the same pattern as for snow cover. This negative correlation was expected based on the more extensive snow cover and frozen ground found further from the equator which negatively impacts performance for both P and ET products as explained above. GRACE measurements are also subject to the effects of the Glacial Isostatic Adjustment (GIA), the redistribution of mass within the Earth resulting from the end of the last ice age (Wahr et al., 1998). While the GIA signal is removed from GRACE TWSA data products, any errors in the GIA models used in this process will result in higher errors in TWSA where the GIA signal is strongest which correlates with higher latitudes.

Dam storage capacity shows a negative correlation with NSE only for product combinations using GPM as a precipitation product and for the TRMM-MOD16 combination. For other combinations, no significant correlations were found. Total runoff relative to dam storage capacity shows a negative correlation for most product combination except for CHIRPS-GLEAM (positive) and TRMM-GLEAM (no significant correlation).

Runoff ratio shows a negative correlation with NSE for 12 of the 45 combinations, and a positive correlation for 24 of the 45. Runoff ratio is computed as the ratio of discharge from GRDC and precipitation from GPM, and the maximum value found

was 42, indicating potentially erroneous data or a strong proportion of discharge originating from storage depletion or inter-basin transfers. Inter-basin transfers in particular would not be represented in our computation of runoff. The runoff ratio was found to be above 1 for 103 stations (out of 937).

A weak negative correlation was found between drainage area and NSE of the RS-derived runoff for 28 of the combinations. The lack of a strong correlation between NSE and catchment area is surprising as the storage change component from GRACE is expected to perform better over larger catchments, particularly because we limited the catchment size here to catchments larger than 10,000 km^2 while GRACE has an inherent spatial resolution of ~300km (90 000 km2) and has been found to produce reliable estimates of storage change for catchments with areas of more than 50,000 km^2 (Biancamaria et al., 2019). Smaller catchments will also be more susceptible to signal leakage from outside the catchment (Dutt Vishwakarma et al., 2016). Catchment size is also expected to influence the applicability of the hypothesis of negligible subsurface fluxes necessary for the application of Eq. 2 as this hypothesis has been shown to be incorrect for smaller catchments (Bouaziz et al., 2018; Fan and Schaller, 2009). Sahoo et al. (2011) and Lehmann et al. (2022) similarly found no correlation between basin area and water balance closure though their studies were limited to 10 very large basins and basins with areas larger than 65 000 km^2 respectively.

Results for the two discrete variables (dominant land cover type and dominant climate class) are shown in Table 4, Table 5, Table 6 and Table 7.

Variability was found between the results for different land cover types. Results for basins with dominant LU codes 40 and 50 (both types of broadleaved forests, see Table 4) perform better than other land cover types, with median NSE values of 0.21 and 0.14 respectively.

Some land cover classes, for example *Open (15-40%) needleleaved deciduous or evergreen forest (>5m)* (class 90), perform particularly poorly, which can be expected as these have a near complete overlap with higher latitude areas. MOD16 performs better than other products in this LC class with a median NSE value of -0.1 while combinations using the other ET products produces median NSE values between -0.33 and -0.96 (Table 5).

Variability is also observed between climate zones, with tropical (median NSE=.11, and median NSE for tropical monsoon .28, see Table 6 and Appendix ~~1~~ A for the detailed results per climate zone) and temperate zones (median NSE=.08) performing better than arid (median NSE=-.04) and continental zones (median NSE= -.08). The SSEBop and CMRSET products produce the highest NSE values in tropical climates, with median NSE values of 0.17, followed by SEBS at 0.15 (Table 7). In temperate zones, using GPM produces the highest median NSE values of 0.11. Lehmann et al. (2022) also analyzed the water balance closure by climate zone and found that errors were relatively consistent within zones with some exceptions. As in this study, the best performance was observed in the "equatorial rain forest/monsoon" zone. This result is also in agreement with the influence of seasonality of rainfall discussed above and observed by Lorenz et al. (2014). Sahoo et al. (2011) on the other hand did not find consistent behavior based on climate zone.

Table 4: NSE values for basins classified by dominant land cover class (LCC) and percentage of time series with positive NSE, percentage NSE above 0.5, total number of time series with the corresponding land cover (n. series) and corresponding number of catchments (n. catchments)

LCC	Land Cover description GlobCover	Median NSE	% NSE >0	% NSE >0.5	n. series	n. catchments
14	Rainfed croplands	-0.01	45	1	1920	65
20	Mosaic cropland (50-70%) / vegetation (grassland/shrubland/forest) (20-50%)	-0.03	44	1	1080	33
30	Mosaic vegetation (grassland/shrubland/forest) (50-70%) / cropland (20-50%)	0.01	55	0	2220	56
40	Closed to open (>15%) broadleaved evergreen or semi-deciduous forest (>5m)	0.21	75	19	3612	83
50	Closed (>40%) broadleaved deciduous forest (>5m)	0.14	68	4	6045	162
60	Open (15-40%) broadleaved deciduous forest/woodland (>5m)	-0.12	36	0	417	19
70	Closed (>40%) needleleaved evergreen forest (>5m)	-0.21	25	1	2619	62
90	Open (15-40%) needleleaved deciduous or evergreen forest (>5m)	-0.61	16	1	2547	173
100	Closed to open (>15%) mixed broadleaved and needleleaved forest (>5m)	-0.53	5	0	390	17
110	Mosaic forest or shrubland (50-70%) / grassland (20-50%)	-2.51	0	0	30	2
120	Mosaic grassland (50-70%) / forest or shrubland (20-50%)	-0.04	28	0	297	8
130	Closed to open (>15%) (broadleaved or needleleaved, evergreen or deciduous) shrubland (<5m)	-0.02	39	2	4086	95
140	Closed to open (>15%) herbaceous vegetation (grassland, savannas or lichens/mosses)	-0.03	34	0	4557	116
150	Sparse (<15%) vegetation	-0.72	23	0	1521	75
180	Closed to open (>15%) grassland or woody vegetation on regularly flooded or waterlogged soil - Fresh, brackish or saline water	-0.07	17	0	36	1
200	Bare areas	-0.33	11	0	297	7
210	Water bodies	-0.48	0	0	60	4

Table 5: Median NSE values per product and per dominant LU class. Cells in italic bold have median values>0, and cells in bold >0.1. Empty cells represent a category where a specific product is not available.

TRMM	GPM	CHIRPS	SSEBOP	MOD16	SEBS	GLEAM	CMRSET	JPL	CSR	GFZ

Dominant Land Cover Class	Med. NSE	Med. NSE	Med. NSE	Med. NSE	Med. NSE	Med. NSE	Med. NSE	Med. NSE	Med. NSE	Med. NSE	Med. NSE
14	***0.01***	-0.04	-0.02	-0.14	-0.03	-0.01	-0.0	0.01	-0.01	-0.01	-0.02
20	***0.02***	-0.11	-0.03	***0.04***	-0.32	***0.04***	-0.11	-0.06	-0.03	-0.03	-0.03
30	***0.01***	***0.02***	***0.01***	-0.01	-0.0	***0.04***	***0.03***	***0.02***	***0.01***	***0.01***	***0.01***
40	**0.21**	**0.18**	**0.24**	**0.21**	**0.21**	**0.25**	***0.07***	**0.27**	**0.2**	**0.2**	**0.21**
50	**0.14**	**0.17**	**0.12**	**0.19**	**0.19**	**0.12**	***0.05***	**0.1**	**0.14**	**0.15**	**0.14**
60	-0.07	-0.13	-0.12	0.07	-0.46	-0.13	-0.13	0.01	-0.13	-0.11	-0.12
70	-0.25	-0.17	-0.2	-0.12	***0.02***	-0.34	-0.17	-1.43	-0.22	-0.22	-0.2
90	-0.42	-0.62	-0.41	-0.58	-0.1	-0.33	-1.12	-0.96	-0.61	-0.6	-0.62
100	-0.53	-0.47	-0.78	-0.41	-0.21	-0.52	-0.79	-1.1	-0.53	-0.55	-0.5
110	-	-2.51	-	-3.72	-7.14	-2.66	-4.11	-2.41	-2.45	-2.6	-2.7
120	-0.03	-0.08	-0.02	-0.29	-0.5	***0.0***	-0.04	-0.03	-0.04	-0.04	-0.04
130	-0.02	-0.03	-0.02	-0.03	-0.01	-0.0	-0.01	-0.09	-0.02	-0.02	-0.02
140	-0.03	-0.04	-0.01	-0.04	-0.05	-0.0	-0.02	-0.02	-0.02	-0.03	-0.03
150	-0.18	-0.88	-0.32	-0.84	-0.27	-0.44	-1.33	-0.79	-0.72	-0.72	-0.69
180	-0.09	-0.08	***0.0***	-0.02	-0.66	-	-0.1	-0.06	-0.06	-0.07	-0.07
200	-0.29	-0.35	-0.32	-0.06	-0.27	-0.19	-0.33	-2.01	-0.34	-0.32	-0.32
210	-0.66	-0.48	-0.8	-0.22	-0.25	-	-0.7	-1.24	-0.5	-0.49	-0.47

Table 6: NSE values for basins classified by climate class

Climate class		Median NSE	% NSE >0	% NSE >0.5	n. series	n. catchments
A	**Tropical**	0.11	67	12	5301	127
B	**Arid**	-0.04	30	1	6483	153
C	**Temperate**	0.08	63	3	6039	162
D	**Continental**	-0.08	35	1	13509	526
E	**Polar**	0.02	52	0	402	11

Table 7: Median NSE values per product and dominant climate class. Cells in italic bold have median values>0, and cells in bold >0.1.

		TRMM Med. NSE	GPM Med. NSE	CHIRPS Med. NSE	SSEBOP Med. NSE	MOD16 Med. NSE	SEBS Med. NSE	GLEAM Med. NSE	CMRSET Med. NSE	JPL Med. NSE	CSR Med. NSE	GFZ Med. NSE
A	**Tropical**	**0.12**	**0.11**	**0.12**	**0.17**	***0.03***	**0.15**	***0.02***	**0.17**	**0.11**	**0.12**	**0.12**
B	**Arid**	-0.03	-0.05	-0.04	-0.05	-0.07	-0.01	-0.03	-0.12	-0.04	-0.04	-0.04

C	**Temperate**	***0.05***	**0.11**	***0.08***	***0.07***	***0.08***	**0.1**	***0.06***	***0.07***	***0.08***	***0.08***	***0.08***
D	**Continental**	-0.03	-0.15	-0.04	-0.09	0.02	-0.11	-0.12	-0.19	-0.08	-0.08	-0.08
E	**Polar**	**0.16**	-0.0	-0.03	**0.34**	**0.13**	**0.28**	-0.32	-0.28	***0.01***	***0.01***	***0.02***

3.4 Results considering anomalies

Remote sensing products are known to be subject to biases and in the results presented so far, no bias correction was considered. In order to investigate how biases may impact the results, we computed the NSE using the anomalies from the mean of the computed runoff and GRDC data. The anomalies from the mean were computed by subtracting the mean of each time series from the time series values.

Considering anomalies rather than absolute values produces a shift in the distribution of the computed NSE values towards higher values (Figure 8) with the percentage of timeseries reaching NSE>0 going from 44.9% to 72.1%, and the percentage of timeseries reaching NSE>0.5 from 3.4% to 4.8%.

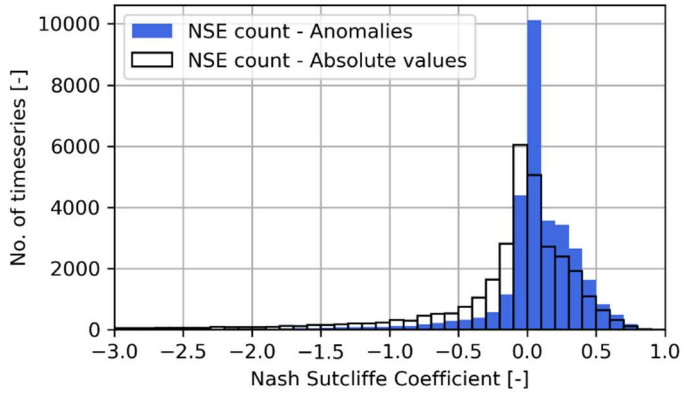

Figure 8: Distribution of NSE values for all time-series for the standard and anomaly time-series. (Time series with NSE<-3 not shown: 2.8% of timeseries for standard and 0.7% for the anomaly time-series)

Increases in NSE for the anomaly time series are most pronounced in the areas which had very low NSE values (see Figure 3 and Figure 10), but many of these retain low NSE values as can be seen for example in the north-western Americas in Figure 9Figure 9.

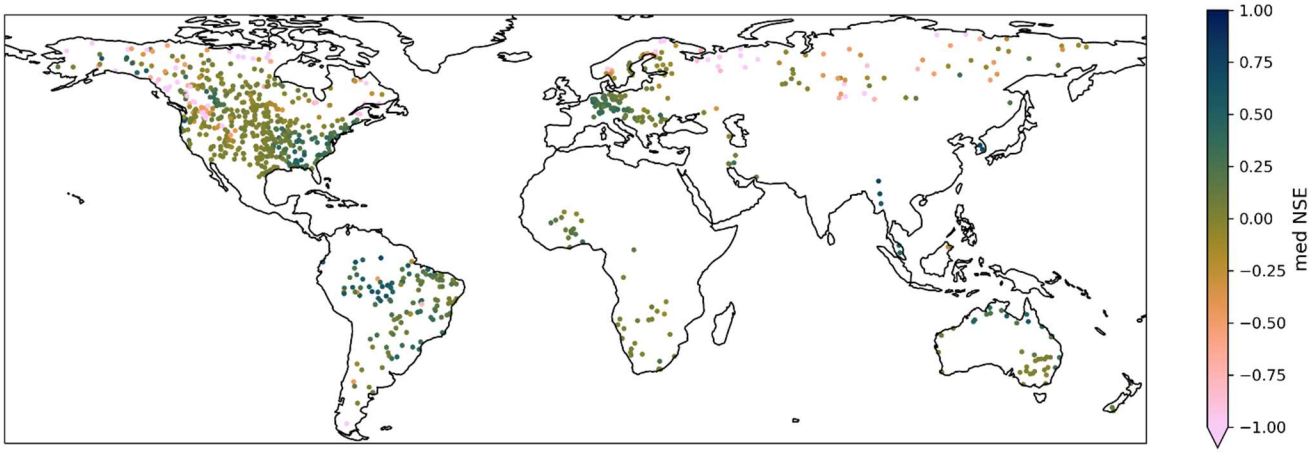

Figure 9: Median NSE for the anomaly time-series for different product combinations at each GRDC station. 49 stations have a median NSE of below -1, the color scale was cropped for legibility.

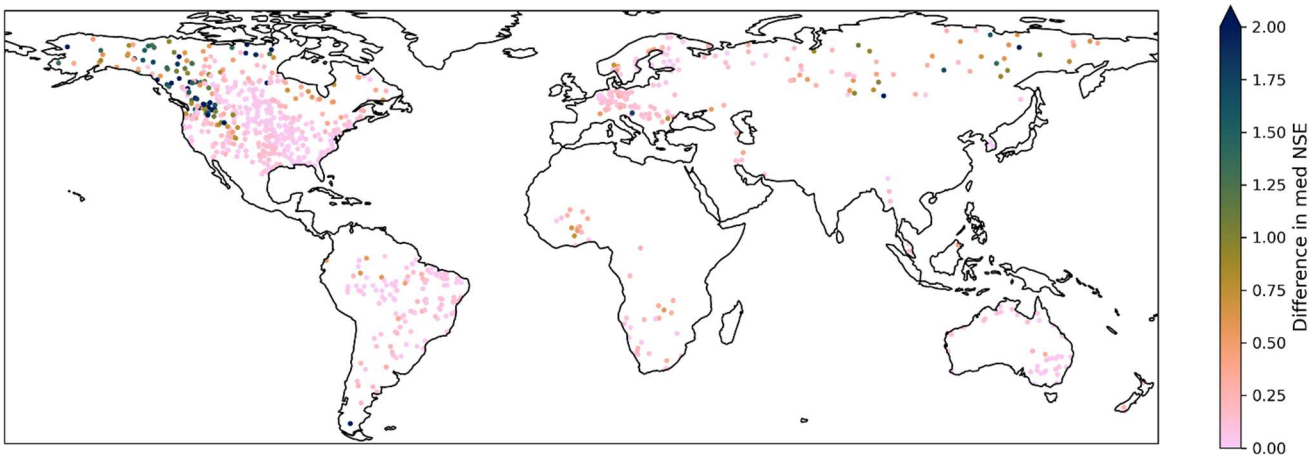

Figure 10: Difference in median NSE values between anomaly and original timeseries. Positive values denote an increase in NSE for anomaly time series – all stations saw an increase in median NSE by moving to the anomaly. 26 stations see an increase of more than 2, the color scale was cropped for legibility.

Results in terms of correlation of NSE with catchment characteristics show some differences in the magnitude of the correlations but very few in the sign of the correlation with the notable exception of the correlations between runoff to precipitation ratio for GPM products. We therefore expect that while using NSE for the anomalies from the mean may show

some differences, the general conclusions would be similar to those presented for the standard time-series. The table of correlations for the anomaly time-series is shown in Appendix B.

4 Conclusions and perspectives

In this study, we analyzed the closure of the water balance at the monthly time-scale for catchments of more than 10 000km^2 by using remote sensing to compute runoff and comparing the computed runoff to in-situ measurements of discharge from the GRDC using the Nash-Sutcliffe Efficiency as the performance metric. We computed the results for 45 different remote sensing product combinations at between 595 to 931 gauging stations depending on the product combinations and analyzed the results through the lens of both the remote sensing products and of 10 catchment characteristics which we computed globally.

Overall, a positive NSE could be reached for at least one product combination for 73.7% of the stations considered. While some product combinations showed better results than others, no one combination or product stood out as systematically performing better than the others. Correlations were found between the NSE values obtained and the ability of remote sensing to close the water balance between areas with different precipitation patterns, in areas with large snow-cover, in different climatic zones and in areas with different dominant land cover classes. This highlights the importance of validating RS products widely. In particular, our results point to the necessity of the improvement of products in continental and arid climate zones and some land covers.

While a number of catchments characteristics were analyzed, these are not exhaustive and for those chosen could have also been computed differently. For example, for larger basins, selecting only one land use category as representative can obscure some differences, and using percentages of area under different types of vegetation may help to further refine results. The same may be considered for climate class. Some additional characteristics which could be interesting to investigate are percentage of area under irrigation in particular for potentially differentiating the different ET products and as a measure of the degree of alteration. One limitation for such an analysis would be the accuracy of global irrigation maps. Some examples of other catchment characteristics which suffer from similar limitations in terms of global data availability or quality but would be of interest are soil type and hydrogeology.

Many satellite products are also calibrated in specific areas though it is not always straightforward to obtain this information consistently. It would be very interesting to assess how different the performance is in areas where calibration activities are carried out versus others and how this impacts the choice of product. These areas could also be correlated with areas with a high density of GRDC stations. Efforts to collect discharge data in underrepresented areas should be undertaken to be included in future studies.

Code and data availability

All datasets used for this study are freely available online, refer to their respective publications to find more details. Code written and used by Bert Coerver and Claire Michailovsky to process the datasets and create the graphs and figures shown in this study is available at https://doi.org/10.5281/zenodo.8318720.

Author contribution

The study was designed by Claire Michailovsky and Bert Coerver. The code to process and analyze data was developed and run by Bert Coerver and Claire Michailovsky. The article was written by Claire Michailovsky and Bert Coerver with input from all the co-authors.

Competing interests

The author Graham Jewitt is a member of the editorial board of HESS. The peer-review process was guided by an independent editor, and the authors have no other competing interests to declare.

Acknowledgements

The authors wish to thank Bich Tran for productive scientific discussions, as well as Dr. Roelof Rietbroek and 3 anonymous reviewers for their comments which have improved the paper.

This research was supported by the Water and Development Partnership Programme (DUPC2) of IHE Delft Institute for Water Education, funded by the Dutch ministry of Foreign Affairs.

References

Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A. and Wood, E. F.: Present and future köppen-geiger climate classification maps at 1-km resolution, Sci. Data, 5(1), 1–12, doi:10.1038/sdata.2018.214, 2018.

Behrangi, A., Gardner, A., Reager, J. T., Fisher, J. B., Yang, D., Huffman, G. J. and Adler, R. F.: Using GRACE to Estitmate Snowfall Accumulation and Assess Gauge Undercatch Corrections in High Latitudes, J. Clim., 31(21), 8689–8704, doi:10.1175/JCLI-D-18-0163.1, 2018.

Biancamaria, S., Mballo, M., Le Moigne, P., Sánchez Pérez, J. M., Espitalier-Noël, G., Grusson, Y., Cakir, R., Häfliger, V., Barathieu, F., Trasmonte, M., Boone, A., Martin, E. and Sauvage, S.: Total water storage variability from GRACE mission and hydrological models for a 50,000 km2 temperate watershed: the Garonne River basin (France), J. Hydrol. Reg. Stud., 24, 100609, doi:10.1016/j.ejrh.2019.100609, 2019.

Bouaziz, L., Weerts, A., Schellekens, J., Sprokkereef, E., Stam, J., Savenije, H. and Hrachowitz, M.: Redressing the balance: Quantifying net intercatchment groundwater flows, Hydrol. Earth Syst. Sci., 22(12), 6415–6434, doi:10.5194/HESS-22-6415-2018, 2018.

Chen, X., Su, Z., Ma, Y., Trigo, I. and Gentine, P.: Remote Sensing of Global Daily Evapotranspiration based on a Surface Energy Balance Method and Reanalysis Data, J. Geophys. Res. Atmos., 126(16), e2020JD032873, doi:10.1029/2020JD032873, 2021.

Dutt Vishwakarma, B., Devaraju, B. and Sneeuw, N.: Minimizing the effects of filtering on catchment scale GRACE solutions, Water Resour. Res., 52(8), 5868–5890, doi:10.1002/2016WR018960, 2016.

ESA and UCLouvain: GlobCover2009, [online] Available from: http://due.esrin.esa.int/page_globcover.php, 2010.

Fan, Y. and Schaller, M. F.: River basins as groundwater exporters and importers: Implications for water cycle and climate modeling, J. Geophys. Res. Atmos., 114(D4), 4103, doi:10.1029/2008JD010636, 2009.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A. and Michaelsen, J.: The climate hazards infrared precipitation with stations - A new environmental record for monitoring extremes, Sci. Data, 2(1), 1–21, doi:10.1038/sdata.2015.66, 2015.

GRDC: Watershed Boundaries of GRDC Stations / Global Runoff Data Centre, 2011.

GRDC: The global runoff data centre, 56068 Koblenz, Germany., 2019.

Guerschman, J. P., Van Dijk, A. I. J. M., Mattersdorf, G., Beringer, J., Hutley, L. B., Leuning, R., Pipunic, R. C. and Sherman, B. S.: Scaling of potential evapotranspiration with MODIS data reproduces flux observations and catchment water balance observations across Australia, J. Hydrol., 369(1–2), 107–119, doi:10.1016/J.JHYDROL.2009.02.013, 2009.

Hall, D. K., Riggs, G. A. and Salomonson, V. V.: MODIS/Terra Snow Cover 5-Min L2 Swath 500m, Version 5, , doi:https://doi.org/10.5067/ACYTYZB9BEOS, 2006.

Huffman, G. J.: The Transition in Multi-Satellite Products from TRMM to GPM (TMPA to IMERG) The transition from the Tropical Rainfall Measuring Mission (TRMM) data products to the Global Precipitation Measurement (GPM) mission products is completed. This document specifically addresses the multi-satellite products, namely the TRMM Multi-satellite Precipitation Analysis (TMPA), the real-time TMPA (TMPA-RT), and the Integrated Multi-satellitE Retrievals for GPM (IMERG), , doi:10.1007/978-3-030-24568, 2020.

Huffman, G. J., Adler, R. F., Bolvin, D. T., Gu, G., Nelkin, E. J., Bowman, K. P., Hong, Y., Stocker, E. F. and Wolff, D. B.: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales, J. Hydrometeorol., 8(1), 38–55, doi:10.1175/JHM560.1, 2007.

Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J. and Jackson, T.: GPM IMERG Final Precipitation L3 1 month 0.1 degree x 0.1 degree V06, Greenbelt, MD, Goddard Earth Sci. Data Inf. Serv. Cent., doi:10.5067/GPM/IMERG/3B-MONTH/06, 2019.

Landerer, F. W.: CSR TELLUS GRACE Level-3 Monthly LAND Water-Equivalent-Thickness Surface-Mass Anomaly Release 6.0 in netCDF/ASCII/Geotiff Formats, , doi:10.5067/TELND-3AC06, 2019a.

Landerer, F. W.: GFZ TELLUS GRACE Level-3 Monthly LAND Water-Equivalent-Thickness Surface-Mass Anomaly Release 6.0 in netCDF/ASCII/Geotiff Formats, , doi:10.5067/TELND-3AG06, 2019b.

Landerer, F. W.: JPL TELLUS GRACE Level-3 Monthly LAND Water-Equivalent-Thickness Surface-Mass Anomaly Release 6.0 in netCDF/ASCII/Geotiff Formats, , doi:10.5067/TELND-3AJ06, 2019c.

Landerer, F. W. and Swenson, S. C.: Accuracy of scaled GRACE terrestrial water storage estimates, Water Resour. Res., 48(4), 4531, doi:10.1029/2011WR011453, 2012.

Lehmann, F., Dutt Vishwakarma, B. and Bamber, J.: How well are we able to close the water budget at the global scale?, Hydrol. Earth Syst. Sci, 26, 35–54, doi:10.5194/hess-26-35-2022, 2022.

Lehner, B., Liermann, C. R., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., Döll, P., Endejan, M., Frenken, K., Magome, J., Nilsson, C., Robertson, J. C., Rödel, R., Sindorf, N. and Wisser, D.: High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management, Front. Ecol. Environ., 9(9), 494–502, doi:10.1890/100125, 2011.

Lorenz, C., Kunstmann, H., Devaraju, B., Tourian, M. J., Sneeuw, N. and Riegger, J.: Large-scale runoff from landmasses: A global assessment of the closure of the hydrological and atmospheric water balances, J. Hydrometeorol., 15(6), 2111–2139, doi:10.1175/JHM-D-13-0157.1, 2014.

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

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D. and Veith, T. L.: Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations, Trans. ASABE, 50(3), 885–900, doi:10.13031/2013.23153, 2007.

Mu, Q., Zhao, M. and Running, S. W.: Improvements to a MODIS global terrestrial evapotranspiration algorithm, Remote Sens. Environ., 115(8), 1781–1800, doi:10.1016/j.rse.2011.02.019, 2011.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I - A discussion of principles, J. Hydrol., 10(3), 282–290, doi:10.1016/0022-1694(70)90255-6, 1970.

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 Sens. Environ., 115(8), 1850–1865, doi:10.1016/j.rse.2011.03.009, 2011.

Senay, G. B., Bohms, S., Singh, R. K., Gowda, P. H., Velpuri, N. M., Alemu, H. and Verdin, J. P.: Operational Evapotranspiration Mapping Using Remote Sensing and Weather Datasets: A New Parameterization for the SSEB Approach, JAWRA J. Am. Water Resour. Assoc., 49(3), 577–591, doi:10.1111/jawr.12057, 2013.

Sheffield, J., Wood, E. F., Pan, M., Beck, H., Coccia, G., Serrat-Capdevila, A. and Verbist, K.: Satellite Remote Sensing for Water Resources Management: Potential for Supporting Sustainable Development in Data-Poor Regions, Water Resour. Res., 54(12), 9724–9758, doi:10.1029/2017WR022437, 2018.

Swenson, S. and Wahr, J.: Estimating Large-Scale Precipitation Minus Evapotranspiration from GRACE Satellite Gravity Measurements, J. Hydrometeorol., 7(2), 252–270, doi:10.1175/JHM478.1, 2006.

Tang, G., Clark, M. P., Papalexiou, S. M., Ma, Z. and Hong, Y.: Have satellite precipitation products improved over last two decades? A comprehensive comparison of GPM IMERG with nine satellite and reanalysis datasets, Remote Sens. Environ., 240, 111697, doi:10.1016/J.RSE.2020.111697, 2020.

Tian, Y. and Peters-Lidard, C. D.: A global map of uncertainties in satellite-based precipitation measurements, Geophys. Res. Lett., 37(24), n/a-n/a, doi:10.1029/2010GL046008, 2010.

Tian, Y., Liu, Y., Arsenault, K. R. and Behrangi, A.: A new approach to satellite-based estimation of precipitation over snow cover, http://dx.doi.org/10.1080/01431161.2014.930208, 35(13), 4940–4951, doi:10.1080/01431161.2014.930208, 2014.

Wahr, J., Molenaar, M. and Bryan, F.: Time variability of the Earth's gravity field: Hydrological and oceanic effects and their possible detection using GRACE, J. Geophys. Res. Solid Earth, 103(B12), 30205–30229, doi:10.1029/98JB02844, 1998.

Wahr, J., Swenson, S., Zlotnicki, V. and Velicogna, I.: Time-variable gravity from GRACE: First results, Geophys. Res. Lett., 31(11), doi:10.1029/2004GL019779, 2004.

Wang, L., Kaban, M. K., Thomas, M., Chen, C. and Ma, X.: The Challenge of Spatial Resolutions for GRACE-Based Estimates Volume Changes of Larger Man-Made Lake: The Case of China's Three Gorges Reservoir in the Yangtze River, Remote Sens., 11(1), 99, doi:10.3390/rs11010099, 2019.

Xu, T., Guo, Z., Xia, Y., Ferreira, V. G., Liu, S., Wang, K., Yao, Y., Zhang, X. and Zhao, C.: Evaluation of twelve evapotranspiration products from machine learning, remote sensing and land surface models over conterminous United States, J. Hydrol., 578, 124105, doi:10.1016/j.jhydrol.2019.124105, 2019.

Zhang, K., Kimball, J. S. and Running, S. W.: A review of remote sensing based actual evapotranspiration estimation, Wiley Interdiscip. Rev. Water, 3(6), 834–853, doi:10.1002/wat2.1168, 2016.

Appendix ~~1~~A: Full result tables for all combinations and by climate zone

Table A1 - Median NSE values for the 45 product combinations. n. NSE>x is the number of time series for which NSE>x and n. catchments is the number of series considered for the specific combination (1 per catchment). The results are presented both for all GRDC stations available for each combination (A.) and for the GRDC stations common to all product combinations (B.)

	median	n.	n.	n.	median	n.	n.	n.
Product Combination	NSE	NSE>0.5	NSE>0	catchmts	NSE	NSE>0.5	NSE>0	catchmts
	A: For all possible catchments				**B: For catchments common to all products**			
TRMM - SSEBOP - JPL	0.0	24	296	601	0.0	24	295	590
TRMM - SSEBOP - GFZ	0.0	24	286	601	0.0	24	285	590
TRMM - SSEBOP - CSR	0.0	25	292	601	0.0	25	290	590
TRMM - CMRSET - JPL	0.0	21	291	604	0.0	21	288	590
TRMM - CMRSET - GFZ	0.0	22	290	604	0.0	22	287	590
TRMM - CMRSET - CSR	0.0	21	292	604	0.0	21	289	590
TRMM - GLEAM - JPL	-0.01	10	259	599	-0.01	10	256	590
TRMM - GLEAM - GFZ	-0.01	10	253	599	-0.01	10	250	590
TRMM - GLEAM - CSR	-0.01	10	256	599	-0.01	10	253	590
TRMM - SEBS - JPL	0.01	25	324	602	0.01	25	320	590
TRMM - SEBS - GFZ	0.01	26	326	602	0.01	26	321	590
TRMM - SEBS - CSR	0.01	25	322	602	0.01	25	317	590
TRMM - MOD16 - JPL	0.0	14	310	595	0.0	14	309	590
TRMM - MOD16 - GFZ	0.0	14	311	595	0.0	14	310	590
TRMM - MOD16 - CSR	0.0	14	308	595	0.01	14	307	590
CHIRPS - SSEBOP - JPL	0.0	35	289	601	0.0	35	286	590
CHIRPS - SSEBOP - GFZ	0.0	35	290	601	0.0	35	287	590
CHIRPS - SSEBOP - CSR	0.0	35	289	601	0.0	35	286	590
CHIRPS - CMRSET - JPL	-0.04	33	251	598	-0.03	33	250	590
CHIRPS - CMRSET - GFZ	-0.04	35	247	598	-0.03	35	246	590
CHIRPS - CMRSET - CSR	-0.03	33	251	598	-0.03	33	250	590
CHIRPS - GLEAM - JPL	-0.01	11	257	599	-0.01	11	254	590
CHIRPS - GLEAM - GFZ	-0.01	13	252	599	-0.01	13	249	590
CHIRPS - GLEAM - CSR	-0.01	12	247	599	-0.01	12	245	590
CHIRPS - SEBS - JPL	0.01	34	348	596	0.01	34	345	590
CHIRPS - SEBS - GFZ	0.01	31	340	596	0.01	31	337	590
CHIRPS - SEBS - CSR	0.01	32	342	596	0.01	32	339	590
CHIRPS - MOD16 - JPL	0.01	26	320	595	0.01	26	318	590
CHIRPS - MOD16 - GFZ	0.0	27	314	595	0.0	27	312	590

CHIRPS - MOD16 - CSR	0.01	26	315	595	0.01	26	313	590
GPM - SSEBOP - JPL	-0.06	26	337	931	0.0	26	282	590
GPM - SSEBOP - GFZ	-0.05	27	336	931	0.0	26	280	590
GPM - SSEBOP - CSR	-0.05	29	330	931	0.0	28	275	590
GPM - CMRSET - JPL	-0.07	28	379	928	0.02	28	331	590
GPM - CMRSET - GFZ	-0.07	31	377	928	0.02	31	326	590
GPM - CMRSET - CSR	-0.08	28	378	928	0.02	29	328	590
GPM - GLEAM - JPL	-0.06	14	327	929	-0.01	14	257	590
GPM - GLEAM - GFZ	-0.06	12	325	929	-0.02	12	257	590
GPM - GLEAM - CSR	-0.06	15	327	929	-0.01	15	256	590
GPM - SEBS - JPL	-0.22	16	319	919	0.0	29	306	590
GPM - SEBS - GFZ	-0.23	17	319	919	0.0	30	296	590
GPM - SEBS - CSR	-0.22	16	320	919	0.0	30	299	590
GPM - MOD16 - JPL	-0.02	23	425	917	0.01	16	307	590
GPM - MOD16 - GFZ	-0.02	20	423	917	0.01	17	307	590
GPM - MOD16 - CSR	-0.02	24	427	917	0.01	18	304	590

Table A2: Full results by climate zone. % NSE>x is the percentage of time series for which NSE>x and n.series the number of time-series produced for each climate class and n.catchments is the number of catchments located in the different climate classes.

Climate class			**Median NSE**	**% NSE>0**	**% NSE>0.5**	**n. series**	**n. catchments**
1	Af	Tropical rainforest climate	0.14	68	11	450	10
2	Am	Tropical monsoon climate	0.28	69	34	945	21
3	Aw/As	Tropical wet and dry or savanna	0.09	66	7	3906	96
4	BWh	Hot desert climate	-0.06	31	0	579	14
5	BWk	Cold desert climate	-0.2	10	0	315	7
6	BSh	Hot semi-arid climate	-0.01	46	6	1137	28
7	BSk	Cold semi-arid climate	-0.04	27	0	4452	104
8	Csa	Hot-summer Mediterranean climate	-0.04	37	0	90	2
9	Csb	Warm-summer Mediterranean climate	-0.0	49	10	441	10
11	Cwa	Monsoon-influenced humid subtropical climate	0.05	55	15	396	21
12	Cwb	Monsoon-influenced temperate oceanic climate	-0.05	41	0	225	5
14	Cfa	Humid subtropical climate	0.12	68	3	3594	89
15	Cfb	Temperate oceanic climate	0.06	61	1	1293	35

18	Dsb	Mediterranean-influenced warm-summer humid continental climate	-0.81	21	0	345	9
19	Dsc	Mediterranean-influenced subarctic climate	-0.07	27	6	135	7
21	Dwa	Monsoon-influenced hot-summer humid continental climate	0.74	100	100	45	3
22	Dwb	Monsoon-influenced warm-summer humid continental climate	-0.06	16	0	105	3
23	Dwc	Monsoon-influenced subarctic climate	-0.66	18	0	120	8
24	Dwd	Monsoon-influenced extremely cold subarctic climate	-0.7	20	0	30	3
25	Dfa	Hot-summer humid continental climate	0.13	74	3	2295	51
26	Dfb	Warm-summer humid continental	-0.08	32	0	7491	248
27	Dfc	Subarctic climate	-0.71	16	1	2871	189
28	Dfd	Extremely cold subarctic climate	-1.25	4	0	72	5
29	ET	Tundra climate	0.06	55	0	312	9
31	EF	Ice cap climate	-0.09	40	0	90	2

Appendix B: Correlation table for anomaly time series

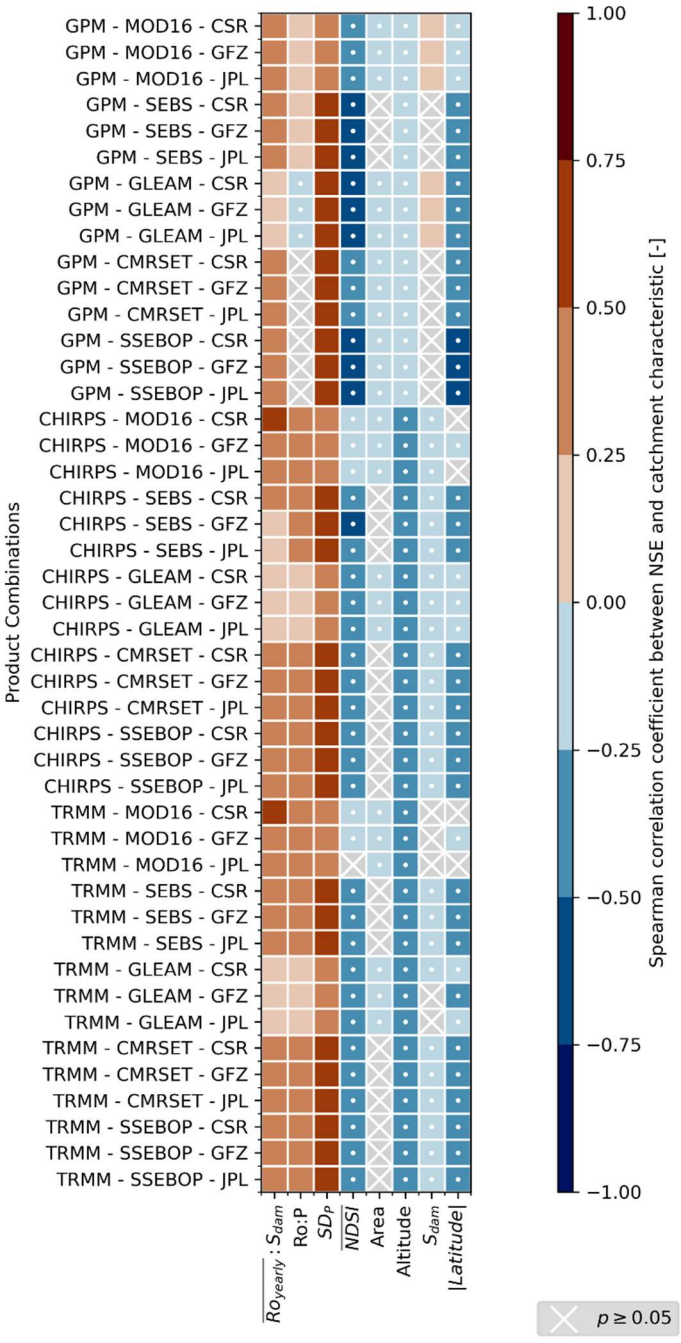

Figure B1: Spearman correlations for different product combinations between the NSEs of anomaly time series for catchments and characteristics of those catchments. See Table 2 for an overview of the catchment characteristics. White dots were added to the negative correlations for monochromatic legibility.