Evaluating the accuracy of gridded water resources reanalysis and evapotranspiration products for assessing water security in ungauged basins

Elias Nkiaka¹, Robert G. Bryant¹, Joshua Ntajal²,³, Eliezer I. Biao⁴

¹Department of Geography, University of Sheffield, Sheffield, S10 2TN, UK
²Department of Geography, University of Bonn, 53115 Bonn, Germany
³Center for Development Research, University of Bonn, 53113 Bonn, Germany
⁴Laboratory of Applied Hydrology, University of Abomey-Calavi (UAC), Cotonou, Benin

Elias Nkiaka (Corresponding author): e.nkiaka@sheffield.ac.uk
Postal Address: Department of Geography, University of Sheffield, Sheffield, S10 2TN, UK

Abstract
Achieving water security in ungauged basins is critically hindered by a lack of in situ river discharge data to assess past, current and future evolution of water resources. To overcome this challenge, there has been a shift toward the use of freely available satellite and reanalysis data products. However, due to inherent bias and uncertainty, these secondary sources require careful evaluation to ascertain their performance before being applied in ungauged basins. The objectives of this study were to evaluate river discharge and evapotranspiration estimates from eight gridded water resources reanalysis (WRR), six satellite-based evapotranspiration (ET) products and ET estimates derived from complimentary relationship (CR-ET) across eight river basins located in Central-West Africa. We also estimated the relative uncertainties in monthly basin-scale water balance evapotranspiration (ETWB) across all the basins. Results highlight strengths and weaknesses of the different WRR in simulating discharge dynamics and ET across the basins. Likewise satellite-based products also show some strength and weaknesses in simulating monthly ET. Analyses further revealed that the relative uncertainties in monthly ETWB range from 4–25 % with a significant increase in magnitude during the rainy season while river discharge appear to be the dominant source of uncertainty. Our results further revealed that the performance of the models in simulating river discharge and evapotranspiration is strongly influenced by model structure, input data and spatial resolution. Considering all the evaluation criteria Noah, Lisflood, AWRAL, and Terra are among the best performing WRR products while Noah, Terra, GLEAM3.5a & 3.5b, and PMLV2 produced ET estimates with the least bias. Considering the plethora of products available, it is imperative to evaluate their performance in representative gauged basins to identify products that can be applied in each region. However, the choice of a particular WRR or ET product will depend on the application and users requirements. Results from this study suggest that gridded WRR and ET products are a useful source of data for assessing water security in ungauged basins.
1. Introduction

River discharge is one of the most important hydrological variables underpinning water resources management, aquatic ecosystems sustainability, flood prediction, and drought warnings at different scales (Mcnally et al., 2017; Couasnon et al., 2020). However, observed river discharge data is often not available at the exact location where critical water management decisions need to be made (Neal et al., 2009). This is especially the case in developing and semi arid/arid regions where discharge gauging stations are sparse (Krabbenhoft et al., 2022), while the number of existing stations are declining (Rodríguez et al., 2020). Despite the acute shortage in observed data, developing regions are areas that are more vulnerable to adverse hydroclimatological conditions (Byers et al., 2018; Kabuya et al., 2020). Furthermore, achieving water security in ungauged basins in developing regions remains a critical development challenge as climate change, population growth, rapid urbanization, and economic growth continue to exert pressure on available water resources under hydrological uncertainty (Flörke et al., 2018; Hirpa et al., 2019). This highlights the urgent need for more reliable data to better assess past, current, and future evolution of water resources, and to predict extreme hydroclimatological events so that better strategies can be put in place to enhance water management and mitigate the impact of extreme events (Nkiaka et al., 2020; Slater et al., 2021). Water security in this study refers to the availability of sufficient quantities of water for human use and ecosystem sustainability.

Evapotranspiration (ET) is another important hydrological variable that represents the linkage between water, energy and carbon cycles and ecosystem services and is the second largest process in the hydrological cycle after precipitation (Zhang et al., 2019). Therefore, ET plays a critical role in water availability at different scales. As such, accurate estimates of ET are also crucial for water management operations such as basin-scale water balance estimation, irrigation planning, estimating water footprint, and assessing the impact of climate change on water availability. However, globally, in situ ET monitoring stations are also scarce while the existing monitoring network cannot provide sufficient information on the temporal and spatial trends of ET at large scales (Laipelt et al., 2021). ET data scarcity may therefore limit our ability to understand changes in the hydrological cycle and water security in the context of environmental change and hydrological uncertainty.

To enhance water security in ungauged basins, there has been a progressive shift toward the use of gridded data derived from satellite and reanalysis (Odusanya et al., 2019; Nkiaka, 2022). This is because gridded data products can provide high spatial resolution and long-term homogeneous data for previously unmonitored areas at scales that are suitable for studying changes in the hydrological cycle and for water management applications (Sheffield et al., 2018). Several gridded data products with global coverage have been produced in recent decades including reanalysis and satellite-based products. Examples of reanalysis products include Watch Forcing Data applied to ERA-Interim
There is also a plethora of satellite products for different hydrometeorological variables such as precipitation, temperature, soil moisture, and ET. For satellite derived ET estimates, it is worth noting that this variable cannot be directly measured by satellites, but rather derived from physical variables observed by satellites from space such as radiation flux. As such, satellite derived ET estimates could rather be referred to as model outputs constrained by satellite data. Another technique used to produce ET estimates is the complimentary relationship (Ma et al., 2021). Considering the way gridded ET products are derived, they tend to suffer from large biases (Weerasinghe et al., 2020; McNamara et al., 2021) and therefore need to be validated before use. In fact, it is argued that validating gridded ET products is an essential step in understanding their applicability and usefulness in water management operations (Blatchford et al., 2020).

Previously, much attention in the development of gridded environmental data was focused on hydrometeorological variables such as precipitation and temperature. However, rapid advancement in computer technology has led to the development of gridded water resources reanalysis (WRR) with quasi global coverage using both land surface models (LSMs) and Global Hydrological Models (GHMs) driven by satellite and reanalysis data. Examples of WRR products include the Global Land Data Assimilation System [GLDAS] (Rodell et al., 2004), “The Global Earth Observation for Integrated Water Resources Assessment” [eartH2Observe] (Schellekens et al., 2017), and the Global Flood Awareness System [GloFAS-ERA5] (Harrigan et al., 2020). Several studies have demonstrated that model-based gridded WRR products can be used as an alternative to observe river discharge in ungauged basins to: (1) understand hydrological processes (Koukoula et al., 2020), (2) support transboundary water management (Sikder et al., 2019), (3) identify flood events (Gründemann et al., 2018; López et al., 2020), and (4) support national water policies (Rodríguez et al., 2020). These examples demonstrate that WRR products have great potential for addressing water security challenges in ungauged basins. Despite their numerous advantages, model outputs from WRR are also fraught with uncertainties resulting from errors in the forcing data, model structure, and the parameterisation of the physical processes in the model scheme (Koukoula et al., 2020). Therefore, it is necessary to evaluate the performance of these products against observed river discharge where available.

Whilst the use of outputs from WRR in water management has gained significant attention in many ungauged or poorly gauged regions such as Asia and Latin America (López et al., 2020; Rodríguez et al., 2020; Sikder et al., 2019), they remain largely under-utilized in Africa. For example, there are only a few case studies reporting on the use of these products in the Upper Blue Nile River basin (Koukoula et al., 2020; Lakew et al., 2020) and the Zambezi River basin (Gründemann et al., 2018). Considering the scale of water insecurity in Africa -compounded by acute data scarcity
(Nkiaka et al., 2021), we feel that evaluating the performance of gridded WRR products in Africa may enhance their adoption in water management in the region. On the other hand, several studies evaluating the performance of gridded data in Africa have focused mostly on precipitation (Dinku et al., 2018; Satgé et al., 2020) while few studies that have evaluated gridded ET products focused on large basins, (Blatchford et al., 2020; Weerasinghe et al., 2020; McNamara et al., 2021) and mostly adopting an annual timescale. This may be attributed to the large scale of the basins which is ideal for the application of satellite data and the coarse spatial resolution of some of the ET products. The availability of high spatial and temporal resolution ET products means that it now possible to evaluate these products in small- to medium-size basins and at a higher temporal resolution. Lastly, considering that the water balance concept has been used widely to evaluate gridded ET products, most studies did not account for uncertainties in basin-wide water balance evapotranspiration (ET\textsubscript{WB}) even though such uncertainties could be large (Baker et al., 2021).

The objectives of this paper were to: (1) evaluate the performance of earth\textsubscript{2}O\textsuperscript{serve} Tier 1 and other WRR products in simulating discharge and evapotranspiration in the basins, (2) evaluate the performance of six satellite-based gridded ET estimates and ET estimates obtained using the complimentary relationship (CR-ET) and (3) estimate the relative uncertainties in ET\textsubscript{WB} in the basins. Considering that only a few studies have attempted to evaluate gridded WRR and ET products over Africa, this paper contributes to the contemporary debate on the performance of these products and how there can be used to assess water security in ungauged basins. We evaluated ET estimates from WRR and other sources considering the fact that users needs for the applicaiton of these products may vary. Hence our evaluation covered a wide range of models and products that meet the needs of different users.

2. Materials and methods
2.1. Study area
The selected basins are located in Central-West Africa ranging in size from 9,000 km\textsuperscript{2} to 499,000 km\textsuperscript{2} (Figure 1). Rainfall in the region is mostly controlled by the north-south movement of the intertropical convergence zone (ITCZ). The main criteria for selecting the basins were: (1) availability of observed river discharge data and (2) for the period of the available discharge data to coincide with the period when gridded WRR and ET data are also available. Additionally, some of the selected basins are facing substantial water security challenges caused by population displacement from conflicts in the Sahel and Lake Chad regions (Kamta et al., 2021; Nagabhatla et al., 2021). The evaluation timestep was determined by the timestep of river discharge data. Shapefiles for all the basins were obtained from HydroSHEDS, locations of the discharge gauging stations were obtained from the respective data sources while the area of each basin was calculated from the basin shapefiles.
HydroSHEDS drainage network offers the unique opportunity to generate watershed boundaries for GRDC gauging stations using a proofed dataset and applying a consistent methodology. Table 1 shows that some of the basins are transboundary in nature.

Table 1: Characteristics of river basins and sources of river discharge data

<table>
<thead>
<tr>
<th>Basin</th>
<th>Total area (km²)</th>
<th>Transboundary (Yes or No) Country(ies)</th>
<th>Population (thousands)</th>
<th>Source of river discharge data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bani</td>
<td>101,600</td>
<td>Yes Ivory Coast, Mali, and Burkina</td>
<td>63,766</td>
<td>GRDC</td>
</tr>
<tr>
<td>Katsina-Ala</td>
<td>22,963</td>
<td>Yes Cameroon and Nigeria</td>
<td>219,875</td>
<td>NHSA</td>
</tr>
<tr>
<td>Konkoure</td>
<td>10,250</td>
<td>No Guinea-Conakry</td>
<td>13,053</td>
<td>GRDC</td>
</tr>
<tr>
<td>Logone</td>
<td>87,953</td>
<td>Yes Cameroon, Chad, and Central Africa Republic</td>
<td>44272</td>
<td>LCBC</td>
</tr>
<tr>
<td>Milo</td>
<td>9,620</td>
<td>No Guinea-Conakry</td>
<td>13,053</td>
<td>GRDC</td>
</tr>
<tr>
<td>Mono</td>
<td>21,575</td>
<td>Yes Togo, Benin</td>
<td>21,479</td>
<td>Co-author</td>
</tr>
<tr>
<td>Oubangui</td>
<td>499,000</td>
<td>Yes Central Africa Republic and the Democratic Republic of Congo</td>
<td>88,742</td>
<td>GRDC</td>
</tr>
<tr>
<td>Oueme</td>
<td>46,990</td>
<td>No Benin</td>
<td>11,488</td>
<td>Co-author</td>
</tr>
</tbody>
</table>

Global River Discharge Centre [GRDC], Nigeria Hydrological Services Agency [NIHSA], Lake Chad Basin Commission [LCBC]. Population data sourced from (Undesa, 2019)
2.2. Input data

2.2.1. Water resources reanalysis (WRR)

The WRR product evaluated in this study include “The Global Earth Observation for Integrated Water Resources Assessment” (eartH2Observe), Famine Early Warning Systems Network [FEWS NET] Land Data Assimilation System (FLDAS), and TerraClimate. The eartH2Observe Tier 1 product consists of a multi-model ensemble of ten global models at a spatial resolution of 0.5° x 0.5° spanning from 1979 to 2012 and driven by Watch Forcing Data methodology applied to ERA-Interim reanalysis (WFDEI) data (Schellekens et al., 2017). The WRR from eartH2Observe project are freely available through the project data portal (https://wci.earth2observe.eu/portal/). Model evaluation here omits the Joint UK Land Environment Simulator (JULES), Simple Water Balance Model (SWBM), and the simple conceptual HBV hydrological model (HBV-SIMREG) as data from the models was not available from the data portal for the selected basins at the time of writing. As such, seven models and model ensemble were included in this study. Evaluation of ET data also omits Lisflood model as data was not available from the portal at the time writing. Although there is an available Tier 2 product with a higher spatial resolution (0.25°), this study did not utilise these data as selected basins were not included at the time of conducting this research. We also evaluated discharge from FLDAS-Noah with spatial resolution of 0.1° and TerraClimate with a spatial resolution of 0.041°. Table 2 provides a brief summary of the different models used in this study.

Table 2: Water resources reanalysis (WRR) products evaluated

<table>
<thead>
<tr>
<th>Model provider</th>
<th>Model name</th>
<th>Model type</th>
<th>Routing scheme</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNRS (Centre National de la Recherche Scientifique)</td>
<td>ORCHIDEE (Organizing Carbon and Hydrology in Dynamic Ecosystems)</td>
<td>LSM</td>
<td>Cascade of linear reservoirs</td>
<td>(Krinner et al., 2005)</td>
</tr>
<tr>
<td>CSIRO (Commonwealth Scientific and Industrial Research Organization)</td>
<td>AWRA-L (Australian Water Resources Assessment)</td>
<td>GHM</td>
<td>Cascade of linear reservoirs</td>
<td>(Van Dijk et al., 2014)</td>
</tr>
<tr>
<td>ECMWF (European Centre for Medium-Range Weather Forecasts)</td>
<td>HITESSEL (Hydrology Tiled ECMWF Scheme for Surface Exchanges over Land)</td>
<td>LSM</td>
<td>CaMa-Flood</td>
<td>(Balsamo et al., 2009)</td>
</tr>
<tr>
<td>JRC (Joint Research Centre)</td>
<td>LISFLOOD</td>
<td>GHM</td>
<td>Double kinematic wave</td>
<td>(Van Der Knijff et al., 2010)</td>
</tr>
<tr>
<td>UniUt (Universiteit Utrecht)</td>
<td>PCR-GLOBWB</td>
<td>GHM</td>
<td>Travel time</td>
<td>(Van Beek et al., 2011)</td>
</tr>
<tr>
<td>MeteoFr (Meteo France)</td>
<td>SURFEX</td>
<td>LSM</td>
<td>TRIP with stream</td>
<td>(Decharme et al., 2010)</td>
</tr>
<tr>
<td>UniK (Universitat Kassel)</td>
<td>WaterGAP</td>
<td>GHM</td>
<td>Manning–Strickler</td>
<td>(Wada et al., 2014)</td>
</tr>
<tr>
<td>NASA</td>
<td>Noah</td>
<td>LSM</td>
<td>Soil-layer water and energy balance</td>
<td>(Mcnally et al., 2017)</td>
</tr>
<tr>
<td>University of California Merced</td>
<td>TerraClimate</td>
<td>GHM</td>
<td>Bucket type model</td>
<td>(Abatzoglou et al., 2018)</td>
</tr>
</tbody>
</table>
2.2.2. Evapotranspiration products

In addition to the ET estimates from the reanalysis products, we also evaluated several satellite-based ET estimates including GLEAM3.5a & 3.5b, MODIS16A2, PMLV1, PMLV2, SSEBop, (see Table 3). ET products from WRR have the same spatial resolution with the discharge estimates while remote sensing products have different spatial resolutions. However, we did not resample the data to the same resolution because a previous study has shown that resampling does not have any significant impact on the results (Weerasinghe et al., 2020). Table 3 provides a summary of all ET products evaluated in this study.

Table 3: Summary of the characteristics of the different ET products

<table>
<thead>
<tr>
<th>ET product</th>
<th>Core equation</th>
<th>Temporal resolution</th>
<th>Spatial resolution</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLEAM3.5a &amp; 3.5b</td>
<td>Priestley-Taylor</td>
<td>Monthly</td>
<td>0.25° x 0.25°</td>
<td>(Martens et al., 2017)</td>
</tr>
<tr>
<td>MODIS16A2</td>
<td>Penman-Montieth</td>
<td>8-day</td>
<td>1/48°x1/48°</td>
<td>(Mu et al., 2007; Mu et al., 2011)</td>
</tr>
<tr>
<td>PMLV1</td>
<td>Penman–Monteith–Leuning</td>
<td>Monthly</td>
<td>0.5° x 0.5°</td>
<td>(Zhang et al., 2016)</td>
</tr>
<tr>
<td>PMLV2</td>
<td>Penman–Monteith–Leuning</td>
<td>8-day</td>
<td>1/192°x1/192°</td>
<td>(Zhang et al., 2019)</td>
</tr>
<tr>
<td>SSEBop</td>
<td>Surface Energy Balance</td>
<td>Monthly</td>
<td>1/96° x 1/96°</td>
<td>(Senay et al., 2013)</td>
</tr>
<tr>
<td>CR-ET</td>
<td>Penman-Montieth</td>
<td>Monthly</td>
<td>0.25°</td>
<td>(Ma et al., 2021)</td>
</tr>
</tbody>
</table>

2.3. Evaluation data

2.3.1. River discharge

Observed river discharge data were used to evaluate the performance of WRR models and to estimate basin-wide water balance evapotranspiration (ET\(_{WB}\)) using the water balance concept. The source of the river discharge data is available in Table 1. Gaps in the discharge data were filled using Self-Organizing Maps which have been shown to be a robust method for infilling missing gaps in hydrometeorological time series (Nkiaka et al., 2016).

2.3.2. Precipitation

Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) was used to estimate ET\(_{WB}\). CHIRPS has a quasi-global coverage at a spatial resolution of 0.05° x 0.05°, spanning the period from 1981 to the present at a daily timescale (Funk et al., 2015). The dataset was designed taking into consideration the weaknesses of existing products (Sulugodu et al., 2019). As such, CHIRPS blends gauge and satellite precipitation covering most global land regions, it has low latency, high resolution, low bias, and long period of record (Funk et al., 2015). CHIRPS has extensively been validated (Dinku et al., 2018; Satgé et al., 2020) and used in several studies in Africa (Larbi et al., 2021; Nkiaka, 2022). The data was downloaded as the spatial average for each basin using the climate engine App and used to estimate ET\(_{WB}\)
2.3.3. GRACE

GRACE data are monthly anomalies of terrestrial water storage changes (TWSC) used to quantify changes in terrestrial water storage. The dataset has a global coverage spanning the period 2003–2017 (Tapley et al., 2019). The data was derived from Jet Propulsion Laboratory (JPL) RL06M Version 2.0 GRACE mascon solution at a spatial resolution of 0.5° x 0.5°. The data has a coastline resolution improvement (CRI) filter to reduce leakage errors across coastlines and land-grids, using scaling factors derived from the community land model (Wiese et al., 2016). GRACE data has recently been re-processed to reduce measurement errors and represents a new generation of gravity solutions that do not require empirical post-processing to remove correlated errors, as such, the present data is better than the previous GRACE version that was based on spherical harmonic gravity solution (Wiese et al., 2016). GRACE data was used in this study to estimate ET\textsubscript{WB} following the approach used in other studies e.g., (Andam-Akorful et al., 2015; Liu, 2018; Xie et al., 2022).

2.4. Evaluating gridded WRR

WRR models were evaluated following a multi-objective approach commonly used in evaluating the performance of hydrological models, including the Nash-Sutcliffe efficiency (NSE), Kling-Gupta efficiency (KGE), and the percent bias (PBIAS). NSE scores range from \(-\infty\) to 1, with 1 indicating a perfect representation of observed discharge. NSE scores \(\geq 0.50\) can be considered acceptable whereas NSE scores \(\leq 0.0\) indicate poor model performance (Moriasi et al., 2007). Similarly, the KGE is a dimensionless metric that can be decomposed into three components crucial for evaluating hydrological model performance accounting for temporal dynamics (correlation), bias errors (observed vs simulated volumes), and variability errors (relative dispersion between observations and simulations) (Gupta et al., 2009). KGE scores range from \(-\infty\) to 1, with 1 considered the ideal value.

Next, PBIAS is used to measure the tendency of the simulated discharge to be larger or smaller than their observed counterparts (Gupta et al., 2009). PBIAS is expected to be 0.0, with low magnitude values indicating accurate simulations, positive values indicate underestimation, negative values indicate overestimation (Moriasi et al., 2007). According to Moriasi et al. (2007), a hydrological model with PBIAS values in the range \(\pm 25\%\) can be considered to be acceptable. Furthermore, a temporal evaluation of flow hydrographs was carried out by plotting the monthly simulated vs observed discharge to ascertain visually if the models were able to capture the magnitude, seasonality, and interannual variability of discharge.

<table>
<thead>
<tr>
<th>Table 4: Contingency table for 80th percentile river discharge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed discharge</td>
</tr>
<tr>
<td>Simulated discharge</td>
</tr>
<tr>
<td>Yes</td>
</tr>
<tr>
<td>No</td>
</tr>
</tbody>
</table>


Lastly, we evaluated the models ability to predict discharge above specific thresholds. This evaluation step is of critical importance when considering operational water management requirements such as water allocation and reservoir operation which rely on monthly river discharge. To achieve this, we adopted the Critical Success Index (CSI) as the metric to evaluate the ability of each model to simulate discharge exceeding the 20th and 80th percentiles. CSI is calculated from a two-dimensional contingency table defining the events in which observed and simulated discharges exceed a given threshold (Thiemig et al., 2015). We used the 20th and 80th percentiles to assess the ability of the models to simulate both low and high flows respectively. The contingency table (Table 4) is a performance measure used in summarizing all possible forecast-observation combinations such as hits (H; event forecasted and observed), misses (M; event observed but not forecasted), false alarms (FA; event forecasted but not observed) and correct negatives (CN; event neither forecasted nor observed). The ideal value for CSI is 100% and the metric is calculated as follows:

\[
CSI = \frac{H}{H + M + FA} \times 100
\]

(1)

2.5. Evaluating gridded ET

We also adopted a multi-step approach to evaluate the performance of ET products by assessing the annual ET–precipitation ratio, evaluating the statistical performance of ET products against long-term ETWB and the ability of the products to capture monthly ET variability.

In the first step, the annual ET–precipitation ratio was calculated to compare with the ratio obtained using ETWB method. The ET–precipitation ratio can also provide an estimate of the amount of water available in each basin after evapotranspiration losses. In the second step, different statistical metrics were used to assess the performance of the ET products using the monthly ETWB as a reference (Andam-Akorful et al., 2015; Burnett et al., 2020; Koukoula et al., 2020). The monthly ETWB was calculated using the basin water balance equation as follows:

\[
ET_{WB} = P - Q - \Delta S
\]

(2)

Where \(P\) is average monthly precipitation over the basin (mm), \(Q\) is river discharge (mm) and \(\Delta S\) is the terrestrial water storage change [TWSC] (mm). Unlike several studies that have evaluated ET products on an annual timescale, this study adopts a monthly sample. As such, the TWSC component (\(\Delta S\)) in equation 2 that is often neglected when estimating ETWB over several years (≥10 years) could not be overlooked. Due to the likely impact of anthropogenic activities such as reservoir operation, water withdrawal, and monthly rainfall variability on TWSC, values derived at monthly timescales are important. In this case, TWSC data used in this study were obtained from GRACE.
Due to the coarse spatial resolution of GRACE, it has been argued that GRACE is not sensitive at detecting changes in monthly TWSC in small-size basins $\leq 150,000$ km$^2$ (Rodell et al., 2011). Based on this claim, it might be argued that GRACE data may not be applicable in this study considering that most of the basins are below this threshold except the Oubangui ($499,000$ km$^2$). However, several studies (Liu, 2018; Biancamaria et al., 2019; Oussou et al., 2022; Xie et al., 2022), have demonstrated that GRACE can provide acceptable TWSC estimates for basins that are smaller than this threshold. Encouraging results from these and other studies do therefore suggest that GRACE data can be used in this study; albeit with the expectation of considerable uncertainties in TWSC estimates. For this study, GRACE data for each basin were obtained by averaging the timeseries of all coincident GRACE grid cells. To estimate changes in monthly TWSC, we calculated the difference between consecutive GRACE measurements for each basin, divided by the time between measurements, using the following equation:

$$\Delta S = (S_{[n]} - S_{[n-1]})/dt$$  \hspace{1cm} (3)

where $\Delta S$ represents the TWSC (mm), $n$ is the measurement number, and $dt$ is the time difference between two consecutive GRACE measurements (months).

Lastly, temporal evaluation of the products was carried out by plotting the time series of all ET products against ET$\_WB$ to visually establish if the gridded ET products were able to capture the magnitude, seasonality, and interannual variability of ET across the basins.

### 2.6. Estimating relative uncertainty in basin-scale water balance ET (ET$\_WB$)

To estimate the relative uncertainty in monthly ET$\_WB$, we first calculated the absolute uncertainty in monthly ET$\_WB$ by propagating errors through each of the components in equation 2 (Rodell et al., 2011), as follows:

$$\sigma_{ET} = \sqrt{\sigma_P^2 + \sigma_Q^2 + \sigma_{\Delta S}^2}$$  \hspace{1cm} (4)

Where $\sigma_P$, $\sigma_Q$ and $\sigma_{\Delta S}$ represent the absolute uncertainties in basin precipitation, observed river discharge, and TWSC respectively. Uncertainty in precipitation was estimated as systematic errors (bias). For this, we used a value of 2% estimated for CHIRPS data at monthly timescale from 1981–2016 over Africa from a validation study using the Global Precipitation Climatology Centre (Shen et al., 2020). Uncertainty in TWSC was determined using the gridded fields of measurement and leakage errors (residual errors after filtering and rescaling) that are provided with the GRACE data. The uncertainty for each basin was calculated by averaging the values of all GRACE grid cells within each basin. To account for month-to-month variation in equation 3, the TWSC error values were
multiplied by $\sqrt{2}$ to obtain $\sigma_{AS}$ (Andam-Akorful et al., 2015). Because no uncertainty estimates were provided with the river discharge data, we adopted a value of 20 % which has been used in a recent study in the region (Burnett et al., 2020). After calculating the absolute uncertainty in monthly ET_{WB}, the relative monthly uncertainty was calculated using equation 5 (Baker et al., 2021) as follows:

$$v_{ET} = \frac{\sigma_{ET}}{ET_{WB}} \times 100$$

Where $v_{ET}$ is the monthly relative uncertainty (%), $\sigma_{ET}$ is the absolute monthly uncertainty (mm), and monthly ET_{WB} (mm). Figure 2 shows a flowchart detailing the different steps used for evaluating the WRR and ET products.

**Figure 2:** Flowchart outlining the steps used in evaluating the WRR and ET products (The blue dotted line in the flow chart separates evaluation of WRR from ET products)

3. Results

3.1. Water resources reanalysis products

3.1.1. Hydrological performance

A multi-objective approach using different statistical metrics (NSE, KGE and PBIAS) was used to evaluate discharge estimates from WRR products. The performance of the models in simulating discharge is shown in Figure 3. Using the NSE as a performance metric, results show that Noah produced positive scores in all the basins (0.15–0.48). Terra, AWRAL and Lisflood produced positive scores (0.01–0.75) in seven, six and four basins respectively. SURFEX model produced positive
scores in three basins while ORCHIDEE, HITESSEL, Watergap and the ensemble mean produced positive scores in two basins each while PCR-GLOBW produced negative scores in all the basins (Figure 3a).

<table>
<thead>
<tr>
<th>Basin</th>
<th>Bani</th>
<th>KatsinaAla</th>
<th>Konkour</th>
<th>Logone</th>
<th>Milo</th>
<th>Mono</th>
<th>Oubangui</th>
<th>Oueme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.51</td>
<td>-0.34</td>
<td>0.13</td>
<td>-4.45</td>
<td>-6.18</td>
<td>-7.21</td>
<td>-1.49</td>
<td>-0.54</td>
</tr>
<tr>
<td>Nash Sutcliff Efficiency (a)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Basin</th>
<th>Bani</th>
<th>KatsinaAla</th>
<th>Konkour</th>
<th>Logone</th>
<th>Milo</th>
<th>Mono</th>
<th>Oubangui</th>
<th>Oueme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.22</td>
<td>-1.23</td>
<td>-0.13</td>
<td>-19.64</td>
<td>-63.95</td>
<td>-24.17</td>
<td>-5.23</td>
<td>-2.36</td>
</tr>
<tr>
<td>Kling, Gupta Efficiency (b)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Basin</th>
<th>Bani</th>
<th>KatsinaAla</th>
<th>Konkour</th>
<th>Logone</th>
<th>Milo</th>
<th>Mono</th>
<th>Oubangui</th>
<th>Oueme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>34.80</td>
<td>112.20</td>
<td>72.80</td>
<td>304.10</td>
<td>58.40</td>
<td>86.30</td>
<td>204.40</td>
<td>104.60</td>
</tr>
<tr>
<td>Percent Bias (c)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3**: Statistical evaluation of the models using (a) NSE, (b) KGE, and (c) PBIAS. Red and orange colours represent poor model performance in Figures 3a, 3b & 3c, however, the acceptable PBIAS range in Figure 3c is ±25%. Ensemble refers to the mean of WRR from the earthH2OObserve.

KGE results show that Noah also produced positive scores (0.11–0.44) in all basins, followed by AWRAL, Lisflood and Terra with positive scores in six, five and four basins respectively (Figure 3b). SURFEX and Watergap produced positive scores in three basins while ORCHIDEE and HITESSEL produced positive scores (0.31–0.76) in two basins. The ensemble mean produced positive scores (0.09–0.42) in three basins while PCR-GLOBW produced the lowest KGE scores (Figure 3b).

Positive and negative PBIAS values were obtained in the different basins. Negative values indicate that the model overestimated discharge volumes compared to observed discharge while positive values indicate the opposite. Noah, Terra and AWRAL produced acceptable PBIAS scores (±25%) in three basins, ORCHIDEE and Watergap produced similar scores in two basins and
HTESSEL in one basin (Figure 3c). The rest of the models including the ensemble mean either grossly overestimated or underestimated discharge volumes in all the basins.

3.1.2. Temporal evaluation

The ability of the models to capture discharge variability was analysed by comparing the simulated vs observed discharge. Results show that most of the models were able to capture the seasonal discharge variability including peak and low flows (Figure 4). However, PCR-GLOBW systematically overestimated low flows and underestimated high flows across all basins. In the Oubangui basin, all models were able to capture the seasonal variability but consistently underestimated peak flows except Noah and Terra models which overestimated peak flows (Figure 4). For example, measured peak discharge in the river exceeds 5000 m³/sec, but all models except Noah and Terra simulated it to be less than 2000 m³/sec (Figure 4).

![Figure 4](image_url): Evaluation of temporal flow variability simulated by the different model

3.1.3. Critical Success Index

Figure 5 shows the performance of the models in simulating the 80th and 20th percentiles monthly discharge. For the 80th percentile flows, results show that Noah and Terra produced CSI scores above 50% in all basins followed by Lisflood and AWRAL in seven and six basins respectively while...
Surfex and Watergap produced similar scores in four basins each (Figure 5a). For the 20<sup>th</sup> percentile flows, only Noah produced CSI scores above 50% in four basins while Lisflood produced similar scores in two basins. The performance of the other models in simulating the 80<sup>th</sup> percentile flow shows a large spread while most models including the ensemble mean failed to simulate the 20<sup>th</sup> percentile flow across all the basins. Taking together, results suggest that the models simulated high flows better than the low flows with only Noah capable of capturing both flow regimes in most basins (Figure 5b).

![Critical Success Index 80th percentile of monthly flow (a)](image)

![Critical Success Index 20th percentile of monthly flow (b)](image)

**Figure 5**: Critical Success Index for 80<sup>th</sup> and 20<sup>th</sup> percentile of monthly flow across all basins

### 3.2. Evapotranspiration products

#### 3.2.1. Evapotranspiration–precipitation ratio

Figure 6 shows the annual ET–precipitation ratio for all basins. It can be observed that average annual ET–precipitation ratio ranges between (0.58–0.76) for WRR and (0.52–0.83) for satellite-based products over a period of 10 years (2003–2012) across all basins. WaterGap produced the highest ratio (0.45-1.01) among WRR models, SSEBop produced the highest ratio (0.53–0.99) while MOD16A2 produced the lowest ratio (0.41–0.66) among the satellite-based products (Figure 6). Results show that the evaporation ratios from the different ET estimates are in the same order of magnitude with the ratio from ET<sub>WB</sub> across all the basins except for WaterGap, SSEBop, MOD16A2 and CR-ET which produced values which were beyond this range (Figure 6).
Figure 6: Annual evapotranspiration – precipitation ratio 2003 – 2012

3.2.2. Basin-wide water balance estimates

Figure 7 shows the results of the statistical metrics used in evaluating the ET estimates using monthly ET$_{WB}$ as reference. Considering bias as a performance metric, AWARL, Noah and Terra produced the lowest bias scores among the estimates from WRR while PMLV2, Terra, and GLEAM3.5a & 3.5b produced the lowest bias scores among the satellite-based products (Figure 7a&d). Most WRR products underestimated ET and similarly GLEAM also slightly underestimated ET, among the satellite-based products while the rest of the products produced mixed results (Figure 7a&d). However, SSEBop systematically overestimated ET in all the basins while MOD16A2 grossly underestimated this variable in all but one basin with respect to monthly ET$_{WB}$ (Figure 7d).
Figure 7: Bias, RMSE, and Pearson correlation coefficient between monthly ET\(_{WB}\) and different ET products (a–c: WRR and d–f: remote sensing products).

Noah produced the lowest RMSE (13–20 mm/month) among the WRR products while GLEAM3.5a & b and PMLV1 produced the lowest RMSE (8.50–12 mm/month) among the satellite-based products (Figure 7b&e). The rest of the products both WRR and satellite-based produced substantially higher RMSE scores (Figure 7b&e). Only Noah and Terra produced high Pearson correlation scores across all basins among WRR products (Figure 7c). On the other hand most satellite-based products produced high Pearson correlation scores (≥0.75) in all basins except PMLV2 and SSEBop which both produced low scores (<0.50) in three and two basins respectively (Figure 7f). ET estimates produced from complimentary relationship (CR-ET) performed poorly across most basins.
**Figure 8a:** Seasonal cycle of ET estimates from WRR and basin-wide water balance evapotranspiration. ET\textsubscript{WB} represents monthly evapotranspiration estimated by the water balance method, while the rest are derived from LSMs and GHMs.

**Figure 8b:** Seasonal cycle of ET estimates from remote sensing-based products and basin-wide water balance evapotranspiration.
3.2.3. Monthly ET variability

Figure 8 shows the seasonal cycle of ET\textsubscript{WB} against both WRR products and satellite-based ET estimates. It can be observed that most products were able to replicate the seasonal ET cycle across all the basins (Figure 8a&b). In addition, most products were not able to replicate the high ET peaks produced by ET\textsubscript{WB} during the rainy season except WaterGap in some instances (Figure 8a). The performance of CR-ET follows that of the rest of the products.

3.2.4. Estimating relative uncertainty in ET\textsubscript{WB}

An assessment of absolute uncertainties in monthly ET\textsubscript{WB} indicated that the dominant sources of uncertainty vary from one basin to another and by each month. For example, in the Katsina-Ala, Konkoure, and Milo basins, the dominant source of uncertainty in monthly ET\textsubscript{WB} was river discharge (supplementary material). Although the absolute uncertainty in precipitation and TWS also appear to be high in the three basins, the uncertainty in river discharge takes precedence over the other sources of uncertainty due to its higher magnitude (supplementary material). On the contrary, the dominant source of uncertainty in ET\textsubscript{WB} in the Bani, Logone, and Oubangui basins was from TWSC. Across all the basins, there was no significant variation in monthly TWSC uncertainty which is consistent with the results of a similar study in the Amazon basin (Baker et al., 2021). Results also revealed that the magnitude of TWSC uncertainty were similar across the basins irrespective of the basin size (Supplementary material).

Figure 9 shows the relative uncertainty in ET\textsubscript{WB} across all the basins. It can be observed that relative uncertainty values are generally <30% but vary from month to month. However, the values were exceptionally high in the Katsina-Ala and Konkoure basins. The relative uncertainty in ET\textsubscript{WB} also appears to be exceptionally high in the months of September–November which corresponds to the high flow season across most basins. Taking together, the average monthly relative uncertainty in ET\textsubscript{WB} for all basins ranges from 10–18% except in the Katsina-Ala and Konkoure basins where this range is grossly exceeded.

![Figure 9: Average (2003 – 2012) monthly relative uncertainty in monthly ET\textsubscript{WB} (%)]
4. Discussion

The overarching goal of this paper was to assess the performance of gridded water resources reanalysis and evapotranspiration products and to estimate the relative uncertainty in monthly basin-wide evapotranspiration ($ET_{WB}$) estimates. Below we provide a discussion and implications of our results in water security assessment in ungauged basins.

4.1. Water resources reanalysis

The performance of WRR products was assessed through commonly used model evaluation metrics, discharge variability, and verification skill scores (critical success index) using observed river discharge data. Our results show strong differences in the performance of the different models in simulating river discharge across the basins. Noah model produced positive NSE and KGE values in all basins and PBIAS values within the acceptable range ($\pm 25\%$) in three basins. Temporal evaluation of the WRR products showed that Noah, Terra, AWRAL and Lisflood were able to capture the seasonal variability in discharge as demonstrated by high KGE scores. Indeed, high KGE values suggest that some models were able to capture the temporal dynamics (strong correlation), and low bias scores indicate that the variability errors between the observed discharge and simulation was also low (Gupta et al., 2009). Nevertheless, Terra consistently overestimated peak flows in all the basins.

Apart from Noah model which is a LSM used in FLDAS, most GHMs used in earthH2Observe tier 1 product performed better than the LSMs, which is consistent with results from other studies (Lakew et al., 2020). The strong performance of GHMs compared to LSMs can be attributed to the differences in the model structure and parametrisation schemes between LSMs and GHMs (Gründemann et al., 2018; Koukoula et al., 2020). For example, some GHMs such as Watergap are able to simulate lakes and reservoirs and water withdrawal while LSMs can only simulate natural processes. Such differences in model structure can significantly influence discharge volumes simulated by both types of models (Gründemann et al., 2018). Although PCRGLOBW is a GHM, it produced substantially low performance compared to the LSMS which is consistent with results from other studies in the region (Gründemann et al., 2018; Lakew et al., 2020). This suggest that PCRGLOBW model may not be suitable for assessing water security in the region.

The ability of the models to simulate flow thresholds was evaluated using the CSI. Results show that Noah, Terra, AWRAL and Lisflood were able to capture more than 50% of 80th percentile monthly flow in most basins. We also noted that apart from Noah model, the rest of the GHMs performed better than the LSMs from earthH2Observe in their ability to capture the 80th percentile monthly flows across the basins while only Noah was able to capture 20th percentile flows in three basins. The performance of Noah compared to other models can be attributed to the fact that FLDAS was specially designed and optimized to produce physically meaningful variables for monitoring food...
and water security in data-scarce regions in Africa (Mcnally et al., 2017). Furthermore, Noah and Terra with spatial resolutions of 0.1° & 0.041° respectively perform better than other models and this may be attributed to their higher spatial resolutions compared to other models with coarser resolution (0.5°). In fact, Gründemann et al. (2018), has shown that WRR products with higher spatial resolution perform better than products with coarser resolution in their ability to simulate discharge. The performance of Noah can also be attributed to the fact the FLDAS is driven by a combination of different precipitation products thereby reducing the uncertainty in the input data while earth2observe tier 1 product are driven by only one data source (WFDEI) which increases the uncertainty in the input data which is propagated to the model outputs. Our results also showed that Lisflood performed better than most other earth2observe models and this can be attributed to the fact that Lisflood has been extensively used in research and operational settings in Africa (Thiemig et al., 2015; Smith et al., 2020). As such, the model parameters may have been better constrained in the region than other models from earth2observe. Taking together, results from this study highlight the importance of evaluating outputs from WRR products in representative basins before applying them in studies that may have wider policy and financial implications. Our results suggest a need to enhance the spatial resolution of WRR products and for the products to be driven by input data from multiple sources to reduce the uncertainties input data.

4.2. Evapotranspiration products

The annual ET–precipitation ratio produced by WRR and satellite-based ET products are within the range estimated for the global land regions (Rodell et al., 2015) with the only exception being WaterGap, SSEBop, MOD16A2 and CR-ET with values beyond this range. This suggests that ET estimates from both sources performed well in this aspect of the ET evaluation. The annual ET–precipitation ratios obtained in this study suggests that annual ET does not exceed annual precipitation in most basins during the period under evaluation. This suggest the availability of sufficient water resources in each basin.

Considering all the ET evaluation criteria and comparing between estimates from WRR and satellite-based products, Noah, Terra, GLEAM3.5a & 3.5b, and PMLV2 appear to outperform the rest of products even though GLEAM products slightly underestimated ET in all the basins. Conversely, WaterGap, SSEBop and MOD16A2 performed poorly and may not be suitable for water security assessment in the region. Our results are generally consistent with those from other studies indicating that GLEAM and MODIS16A2 underestimate evapotranspiration, while SSEBop overestimates this variable in most parts of Africa (Weerasinghe et al., 2020; Adeyeri and Ishola, 2021; Mcnamara et al., 2021). Given that ET estimates from Noah and Terra are produced together with other water balance components (runoff, soil moisture and baseflow) the two models may be
recommended for water security assessment in the region because of water balance closure. Our results also revealed that the performance of satellite-based ET products is not influence by spatial resolution which is consistent with results from previous studies (Weerasinghe et al., 2020; Jiang and Liu, 2021). For example, Gleam products with a spatial resolution of 0.25° outperformed products such as MODIS16A2 and SSEBop with higher spatial resolutions. On the contrary, ET estimates from WRR appear to be influenced by spatial resolution considering that Noah and Terra with higher spatial resolutions perform better than other products with coarser resolutions.

Although all the products were able to capture the temporal dynamics of ET in all the basins, there were substantial differences in the magnitude of monthly ET from each model. This finding is consistent with results from other studies showing strong differences in ET estimates produced by different models (Weerasinghe et al., 2020; Adeyeri and Ishola, 2021). The discrepancies in monthly ET estimates from the models may be attributed to differences model structure, parameters, and uncertainties in the input data used in driving the models. This is also in-line with findings from another study in West Africa highlighting the impact of model parameters and input data uncertainty on ET estimates (Jung et al., 2019). Considering the aforementioned factors, it may be difficult to expect the products to produce similar results. ETWB estimates across all the basins produced high peaks during the rainy season which is also similar to the results of a related study in West Africa (Andam-Akorful et al., 2015). The high peaks observed in ETWB may be attributed to errors inherent in monthly precipitation, river discharge, and TWS estimes used in estimating monthly ETWB.

Given that there was no uncertainty information on the river discharge data used in this study, we adopted a value of 20% following a previous study in the region (Burnett et al., 2020). In fact, we feel that this value may be conservative considering that uncertainties in river discharge in tropical regions have been shown to exceed 200% (Kiang et al., 2018). The mean monthly relative uncertainty for ETWB for most basins appears to be in the same order of magnitude (16%) with results obtained in the Amazon basin (Baker et al., 2021). Results also showed that the relative uncertainty in ETWB is not influenced by basin size as most basins produced similar (same order of magnitude) uncertainty estimates. The relative uncertainty in monthly ETWB was higher during the rainy season. This can be linked to high rainfall input during the rainy season which translates to high river discharge and TWS thereby increasing the absolute uncertainties in the different water balance components used in estimating ETWB. Results from this study suggest that the relative uncertainty in monthly ETWB may be substantial which can potentially influence the performance of ET products when they are evaluated using the ETWB method. We therefore recommend that evaluating the performance of ET products at monthly timescale should be accompanied with the estimation of relative uncertainties.
5. Conclusions

The objectives of this study were to assess the performance of water resources reanalysis and evapotranspiration products and to estimate the relative uncertainties in monthly $ET_{WB}$ across eight basins in Africa. Results show varying strengths and weaknesses for the different models. Some models were able to capture the river discharge dynamics in the basins while other models could not adequately capture this pattern. Differences in the model performance can be attributed to differences in model structure, parameters, input data used in driving the models and the spatial resolution of the WRR products. Apart from Noah which is a land surface model (LSM), global hydrological models (GHMs) performed better than LSMS except PCRGLOBW.

Evaluation of gridded ET products also revealed varying strengths and weaknesses for the different products. Based on the different evaluation criteria (bias, RMSE, Pearson correlation coefficient, and temporal ET variability), Noah appears to outperform most of other ET estimates and may therefore be recommended for water security assessment in the region. More so, because of water balance closure and the availability of other water balance components (runoff, soil moisture and baseflow). Our results also suggest that the performance of satellite-based ET products is not influenced by spatial resolution, while differences in ET estimates may be attributed to differences in model structure, parameters and the input data used to drive each model. On the contrary, spatial resolution appear to have a significant impact on the performance of WRR in simulating ET estimates.

Our results also revealed that the relative uncertainties in monthly $ET_{WB}$ were substantially higher during the rainy season which can be attributed to uncertainties inherent in higher rainfall leading to an increase in discharge magnitude and TWSC during this period. Results also revealed that uncertainty in river discharge is the dominant source of uncertainty in $ET_{WB}$. This underscores the need to prioritize the installation of new gauging stations while upgrading existing stations. This is because uncertainties in river discharge could constrain the ability to fully understand hydrologic variability and undermine discharge prediction.

Results from this study suggest that WRR and ET products may be used for water security assessment in ungauged basins. However, it is imperative to evaluate the performance of these products in representative gauged basins before applying them in ungauged basins. This is because applying the products in ungauged basins without evaluating their performance may lead to poor water management decisions with wider policy and financial implications. However, there is also a need for WRR and ET products to be driven by input data from multiple sources to reduce uncertainties in the input data and at the same time, the spatial resolution of WRR products needs to be enhanced. Results from this study may be used by the products developers to improve on the quality of future generations of WRR and ET products.
**Author contributions:** EN and RGB designed the methodological framework and contributed to the entire strategic and conceptual framework of the study. EN prepared the data, performed the analyses, interpreted the results and wrote the original draft. JN and EIB provided discharge data for the Mono and Oueme basins respectively. All authors read the paper and provided feedback.

**Competing interests:** The authors declare that they have no conflict of interest.

**Acknowledgements:** E.N. was funded by the Leverhulme Trust Early Career Fellowship – Award Number ECF–097–2020. We are grateful to Coralie Adams at Manchester University for writing the Python code that was used to produce Figures 4 & 8.

**References**


Flörke, M., Schneider, C., and McDonald, R. I.: Water competition between cities and agriculture driven by climate change and urban growth, Nature Sustainability, 1, 51-58, https://doi.org/10.1038/s41893-017-0006-8, 2018.


