Reply to Reviewer 3

This paper presents analyses of the water budget and water cycle for Czechia. Overall I found that the work is interesting and well written. My major comment is around the definition of the score used for the ranking of the different data set combinations and how this was derived and justified. For example the score only accounts for the anomalies and the correlations but does not consider bias in the products. This is very evident from Figure 4 where ERA5-land has substantially higher estimates of both P and ET and therefore its anomalies are similar to the other products. But presumably in some applications consistent biases may be problematic even if the anomalies are ok (e.g. water allocations or environmental flows). I think that the authors need to do far more to consider the sensitivity of the dataset ranking to the definition of the score.

We thank the reviewer for their constructive and encouraging comments. As correctly pointed out the score does not account for any biases in the products. However, if precipitation and evapotranspiration are over- or underestimated simultaneously then the overall water budget closure is not significantly affected. The metric proposed herein aims to rank multi-source data combinations to determine how well a given combination of data sets closes the water budget. It is a method that can be used to easily and quickly filter out the data set combinations providing implausible results and then be complemented with additional analyses that consider the bias as we did in the original manuscript. We agree that the approach introduced in our work might not be the best suited for different applications that need to quantify absolute values rather than anomalies in water fluxes. The main aim of our work is not to benchmark the different data sets analyzed herein but to demonstrate how different can become the water cycles depicted by each of them. To clarify this, we will add in the revised manuscript the following:

”Our evaluation of individual water cycle components is cohesive with previous literature. Although the data products assessed herein have been previously analyzed at multiple spatial scales, this is done under a univariate perspective, that does not consider the ability of the data sets to reproduce the water cycle and its changes as a whole in a structurally plausible manner. This is easily denoted by the fact that even though mHM’s performance was the best for all water cycle components evaluated using high-quality observational references, the best data set combination ranking is actually TerraClimate exclusive (i.e., all flux estimates from the same data set). Note that the score metric and ranking framework proposed herein serve as a method that can easily and quickly filter out the data set combinations providing implausible results. It should be remarked that this ranking framework acts as an initial assessment to be complemented with additional analyses because the score metric does not account for any biases in the products. Expressly because our work aims not to benchmark the different data sets analyzed herein but to demonstrate how different can become the water cycles depicted by each of them.”

Minor comments:

Figure 1: shading is difficult to interpret and I think it would be easier to use hatching or just label the rivers

Figure 1 will be revised as suggested:

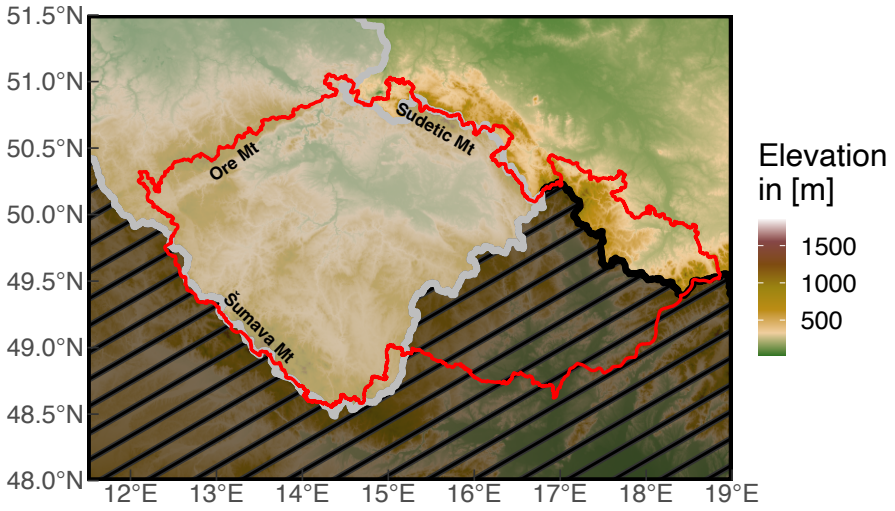

Figure 1. The three drainage basins within Czechia's boundaries. Elbe (light gray shade), Danube (striped dark gray shade), and Oder (no shade).

Line 163: Would be interesting to do the analyses for the three main drainage basins.

The following figure with the corresponding text will be added:

"The water cycle budget is meant to close over hydrological units. Accordingly, we examined the water fluxes of the data sets with the best evaluation over the subbasins enclosed by the Czech administrative borders (Figure 3). For simplicity, we will refer to them as the Danube basin inside Czechia, the Elbe basin inside Czechia, and the Oder basin inside Czechia. It can be seen that within each data set, no extremely deviant behavior is exhibited between basins or at the country level. In other words, the precipitation time series depicted by TerraClimate for Czechia is similar to the one depicted for the Danube, Elbe, and Oder basins inside Czechia. Comparing data sets, however, it is evident that ERA5-Land is different. At first glance, we evince higher magnitudes for ERA5-Land precipitation and evapotranspiration, yet the residuals do not appear to be that far off from those of mHM or TerraClimate. It is not until we look at the cumulative sum of the residuals that we can distinguish ERA5-Land water budget residuals are nonstationary with a decreasing trend."

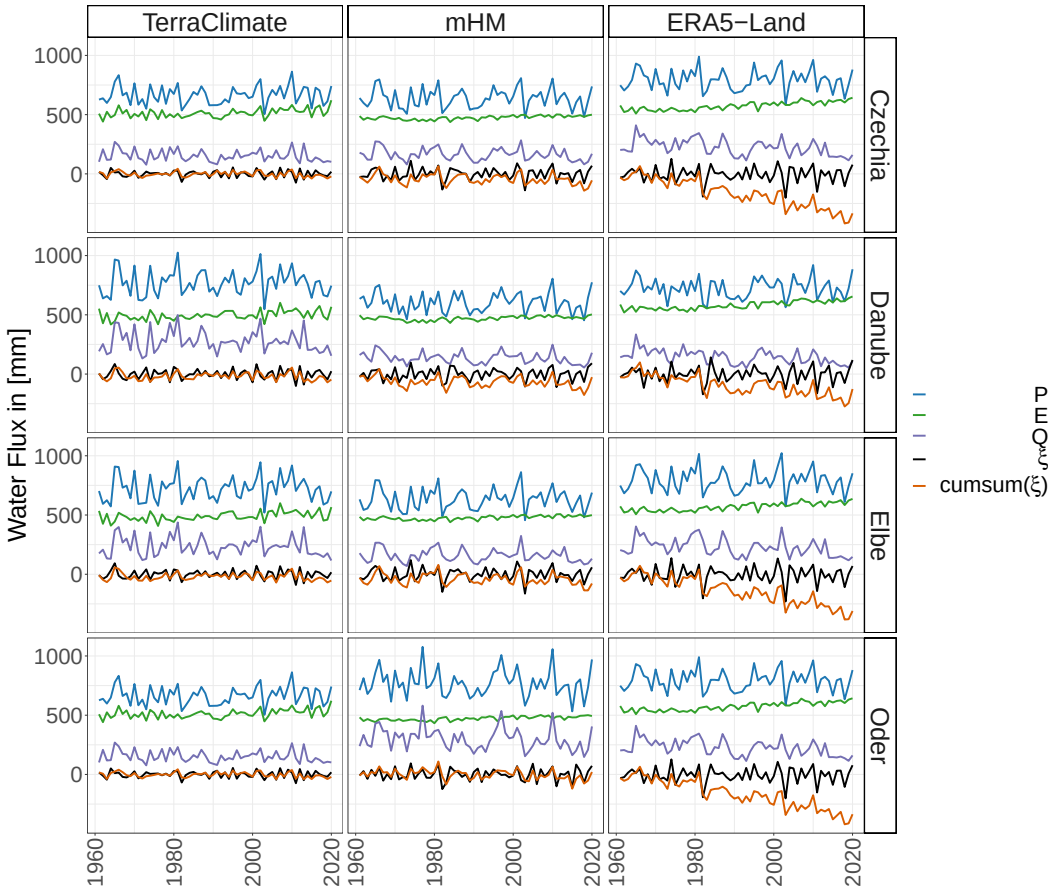

Figure 3. Spatial weighted average annual water fluxes over Czechia (first row), Danube basin inside Czechia (second row), Elbe basin inside Czechia (third row), and Oder basin inside Czechia (fourth row). Where P is precipitation in blue, E is evapotranspiration in green, Q is runoff in purple, ξ is the residual ($P - E - Q$) in black, and cumsum(ξ) is the cumulative sum of the residual in orange. Left column: TerraClimate (P), TerraClimate (E), and TerraClimate (Q). Middle column: mHM(E-OBS) (P), mHM (E), and mHM (Q). Right column: ERA5-Land (P), ERA5-Land (E), and ERA5-Land (Q).

Line 170: would be good to explicitly note that you are doing the closure each year here and then Ri is the average of Rj for j in 1:60

To explicitly describe the average residual we will modify the manuscript from: "A success metric widely used among several studies is getting the budget closure residual (R) as close to zero as possible. Herein, we define the budget closure residual as follows:

$$R = P - E - Q \tag{1}$$

where P is precipitation, E is evapotranspiration, and Q is runoff. Thus, we have 96 distributions of 60 annual values each. The ranking of a given data set combination was determined via:

$$Ranking = \frac{|\overline{R_i}|\sigma_{R_i}}{(cor(P_i - E_i, Q_i)cor(P_i, P_o)cor(E_i, E_o)cor(Q_i, Q_o))^2} \tag{2}$$

where $|\overline{R_i}|$ is the absolute value of the mean of the 60 annual residuals for the i-th combination, σ_{R_i} is the standard deviation of the 60 annual residuals for the i-th combination, $cor(P_i - E_i, Q_i)$

is the correlation between $P-E$ and Q for the i-th combination, $cor(P_i, P_o)$ is the correlation between P of the i-th combination and the precipitation evaluation reference, $cor(E_i, E_o)$ is the correlation between E of the i-th combination and the evapotranspiration evaluation reference, and $cor(Q_i, Q_o)$ is the correlation between Q of the i-th combination and the runoff evaluation reference."

To: "A success metric widely used among several studies is getting the budget closure residual (ξ) as close to zero as possible. Herein, we define the budget closure residual as follows:

$$\xi_n = P_n - E_n - Q_n \tag{3}$$

where P_n is precipitation, E_n is evapotranspiration, and Q_n is runoff for a given year n. Thus, we have 60 annual values for each of the 96 possible combinations. Under steady state conditions the mean of these residuals should tend to zero:

$$\overline{\xi_i} = \frac{\sum_{n=1}^{N} \xi_n}{N} \rightarrow 0 \tag{4}$$

where $\overline{\xi_i}$ is the mean of the $N = 60$ annual residuals for the i-th combination. The score to be used in the ranking of a given data set combination was determined via:

$$score = \frac{|\overline{\xi_i}|\sigma_{\xi_i}}{(cor(P_i - E_i, Q_i)cor(P_i, P_o)cor(E_i, E_o)cor(Q_i, Q_o))^2} \tag{5}$$

where $|\overline{\xi_i}|$ is the absolute value of the mean of the 60 annual residuals for the i-th combination, σ_{ξ_i} is the standard deviation of the 60 annual residuals for the i-th combination, $cor(P_i-E_i, Q_i)$ is the correlation between $P-E$ and Q for the i-th combination, $cor(P_i, P_o)$ is the correlation between P of the i-th combination and the precipitation evaluation reference, $cor(E_i, E_o)$ is the correlation between E of the i-th combination and the evapotranspiration evaluation reference, and $cor(Q_i, Q_o)$ is the correlation between Q of the i-th combination and the runoff evaluation reference."

Equation 2: this isn't actually the ranking but a score that is then used for ranking so I think all the text associated with the equation needs to be updated.

The text will be rephrased from: "The ranking of a given data set combination was determined via:"

To: "The score to be used in the ranking of a given data set combination was determined via:"

Figure 3 - we can't see most of the distributions. I don't think this is a useful presentation of the data. What are the units for the budget residual?

Figure 3 (now Figure 4) will be modified to include only the distributions listed in table 2. The original figure with all the distributions will be placed in the supplementary material as Figure S2.

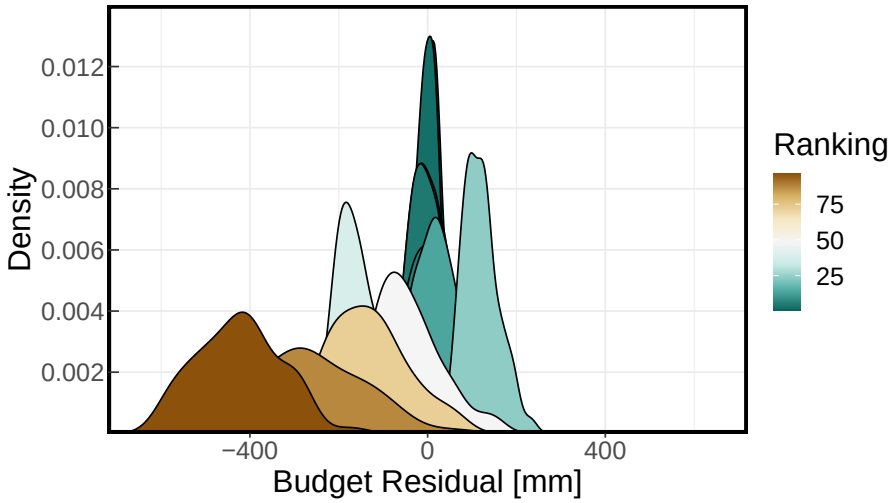

Figure 4. Empirical distribution of the data set combinations listed on 2 colored based on their ranking as determined by Equation 5. The color gradient goes from higher ranked combinations colored in shades green to lower ranked combinations colored in shades of brown.

Figure 5 - wrong colours mentioned in caption. I am surprised by the results shown in figure 5 as there is less difference between the different models than implied by Figure 4 where ERA5 is substantially wetter and higher ET. I think you could dig further into this.

We thank the reviewer for their detailed attention and corresponding suggestions. Captions will be revised to describe the appropriate colors. The story regarding water cycle changes depends on the data set of choice and the time scale. These differences tend to be overlooked when annual averages are being compared, but when it comes to annual totals, the small discrepancies add up, leading to such results. We further highlight some substantial inconsistency in the ERA5-Land data (Figure 3). It appears that the cumulative sum of the water budget residual in ERA5-Land declines monotonically in time, implying some systematic bias in the water budget closure. Even though that approximately 500 mm over 60 years might be considered a relatively small amount, it raises further questions about the applicability of ERA5-Land in hydrological studies and therefore, extra caution should be taken when the widely-used reanalysis data product is employed.