Dear Bob Su,

thank you for inviting us to submit a revised version of the manuscript! In the following, we provide a point-by-point reply to all reviewers' comments. The revised manuscript includes all changes as indicated in the responses, as well as some minor rephrasing to improve clarity.

Author's response to Referee #1

Major Comments

1) **The current manuscript only used two numerical experiments. One with and another without vegetation. And highlight that the VEG is different than the current approach of using plant functional types or land cover classes. Nevertheless, the comparison between the VEG and the 'traditional approach' is not presented. This reviewer suggest the authors to add simulation results of the traditional approach. As such, the add-value of using dynamic vegetation can be demonstrated more clearly.**

AC: This is indeed a valid point that we missed to emphasize in the manuscript. Based on the suggestion, we have not only clarified it in the text, but additionally performed a **PFT** experiment in which we define and calibrate the vegetation-dependent parameters for different plant-functional type (PFT) classes separately and then analyze model performance and TWS composition in comparison to the **B** and **VEG** experiments. The results show that the larger number of parameters (due to different sets for different PFT) does not lead to sizable improvements of model performance, but instead increases parameter uncertainty possibly due to overparameterization. In terms of TWS composition, we see substantial differences in the **PFT** experiment compared to **B** and **VEG,** which underlines our conclusions that the representation of vegetation in GHMs is critical for interpreting TWS variations.

Based on the GSWP2 land cover classification (Dirmeyer et al. 2006), we consider 12 PFT classes (Fig. 1), for which we define individual values of $wSoil_{max(2)}$ (maximum available water capacity of the 2nd soil layer) and s_{berg} (scaling parameter to derive the runoff/infiltration coefficient). Since state-of-the-art global hydrological models (GHMs) usually include seasonal dynamics of leaf area index (LAI) to calculate, e.g., transpiration, we decided to keep the definition of the active vegetation fraction as a function of seasonal EVI data as in the **VEG** experiment. Instead, we focus on $wSoil_{max(2)}$ because GHMs usually apply a PFT specific rooting depth, and on s_{berg} because this is similar to the runoff coefficient γ which is tuned in some GHMs (e.g., the WaterGAP model (Müller Schmied et al. 2021)).

Considering these 12 PFT classes increases the number of calibration parameters from 12 (in **B**) and 16 (in **VEG**) to 34 (in **PFT**). Analysis of parameter uncertainty shows high uncertainties for a set of parameters common with **B**, while optimized parameter values are between those of **B** and **VEG** (Table 1). Additionally, and unlike **B** and **VEG**, **PFT** has high uncertainty of $wSoil_{max(2)}$ for all PFT classes, and high correlation between each PFT's $wSoil_{max(2)}$ and s_{berg} (Fig. 2). High uncertainty of $wSoil_{max(2)}$ is an indication that having one $wSoil_{max(2)}$ per PFT may not explain the within-PFT variability. On the other hand, high correlation between each PFT's $wSoil_{max(2)}$ and s_{berg} is systematic, as both parameters are based on the same spatial distribution of PFT classes - and highlights an advantage of the **VEG** experiment, in which both are based on independent data sets.

In terms of model performance, Fig. 3 shows a partial improvement for wTWS and ET in the **PFT** experiment. Especially in the *Humid* and *Sub-humid* regions, wTWS simulation in **PFT** matches GRACE observations better. They include tropical regions, where data for maximum plant available water capacity by Tian et al. 2019 (RD4), which got the largest weight in the **VEG** experiment, is not available. Note that we filled the missing values for tropical regions with the same $wSoil_{max(RD4)}$ value as in the Northern latitudes. Better performance in the **PFT** experiment suggests a shortcoming of the vegetation implementation in **VEG**, where at least 2 different $wSoil_{max(RD4)}$ fill values seem necessary for different climate regions. In contrast to wTWS and ET, **PFT** performance of Q is poorer than in **B** and **VEG**, with a clear underestimation of the seasonal variability. To consider model performance in relation to the number of calibration parameters, we calculated the Akaike information criterion (AIC). Since low values of AIC indicate better performance compared to the other experiments, **PFT** only performs superior regarding ET, while the increased number of model parameters isn't advantageous regarding wTWS and Q simulations. Also, note that the increased number of model parameters comes at an additional computational cost.

Further, changing the representation of vegetation changes the simulated TWS composition (Fig. 4-6), as the contribution to TWS variability differs between experiments. In **PFT**, among the liquid water storages wSoil contributes most to mean seasonal TWS variability, with Impact Index values between those of **B** and **VEG** (Fig. 4, Fig. 6). Compared to **VEG**, wSlow is in general less important in **PFT**, while wDeep has a less impact on mean seasonal TWS, but it's contribution to inter-annual TWS variability increases.

All in all, this analysis underlines that including continuous fields of vegetation parameters is preferable than the 'traditional' PFT-based approaches of defining parameters for distinct PFT classes (and their calibration) - in terms of model calibration and the uncertainty of calibrated model parameters, but also regarding model performance in relation to the number of model parameters. Further we could highlight that the representation of vegetation in hydrological models is crucial for the partitioning of simulated TWS.

We will include this analysis in detail in the supplement of the revised manuscript, and include the major findings in the discussion of the main text.

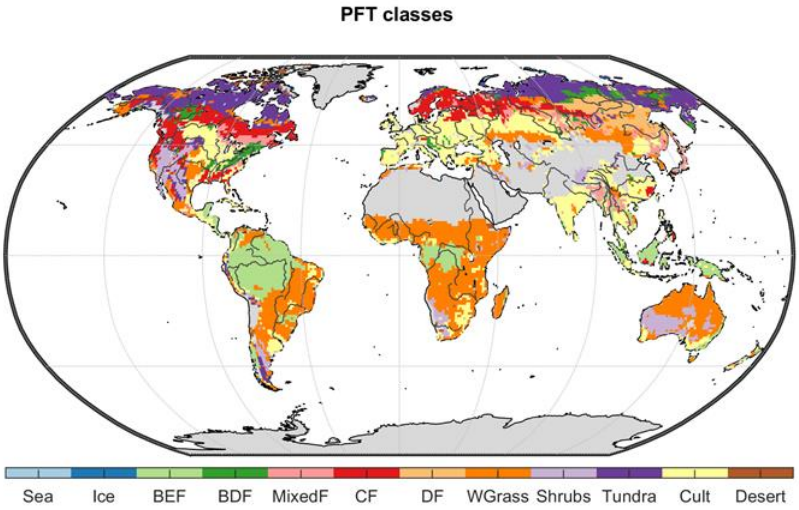

Figure 1: Classes of plant functional type used in the PFT experiment. (Sea (PFT0); Ice=Continental Ice (PFT1); BEF=Broadleaf Evergreen Forest (PFT2); BDF=Broadleaf Deciduous Forest & Woodland (PFT3); MixedF=Mixed Coniferous & Broadleaf Deciduous Forest & Woodland (PFT4); CF= Coniferous Forest & Woodland (PFT5); DF=High Latitude Deciduous Forest & Woodland (PFT6); WGrass= Wooded C4 Grassland (PFT7); Shrubs=Shrubs & Bare Ground (PFT8); Tundra (PFT9); Cult=Cultivation (PFT10); Desert (PFT11)).

Table 1: Calibrated parameter values and their uncertainty for B, VEG and PFT. Red font indicates a calibrated parameter that hits the parameter bounds, and red background indicates parameter uncertainty ≥ 20%.

parameter	calibrated values ± uncertainty					
	B		VEG		PFT	
vegetation fraction						
p_{veg}	0.37	± 0.05				
s_{EVI}			3.89	± 0.05	3.75	± 0.03
evapotranspiration						
p_{int}	1	± 0.08	0.6	± 0.02	0.71	± 0.02
k_{Soil}	0.1	± 0.01	0.4	± 0.08	0.27	± 0.04
α_{veg}	2.25	± 0.15	0.92	± 0.00	0.87	± 0
k_{Transp}	0.12	± 0.32	0.48	± 1.76	0.5	± 4.32
deep soil						
s_{Deep}	9.1	± 461317	5.6	± 0.21	8.48	± 0.24
f_{max}	1.5	± 0.00	5.1	± 0.01	11.77	± 0.02
d_{Deep}	1	± 5.61	0.01	± 0.00	0.03	± 0
delayed water storage						
rf_{Slow}	0.78	± 1.72	0.68	± 0.01	0.62	± 0.05
d_{Slow}	1	± 2329	0.02	± 0.03	0.03	± 0.19
infiltration/runoff						
p_{berg}	1.32	± 0.02				
s_{berg}			3.08	± 0.02		
$s_{berg\_PFT_0}$					3.7	± 0.45
$s_{berg\_PFT_1}$					3.11	± 0.32
$s_{berg\_PFT_2}$					1.87	± 0.01
$s_{berg\_PFT_3}$					2.57	± 0.09
$s_{berg\_PFT_4}$					2.04	± 0.03
$s_{berg\_PFT_5}$					4.31	± 0.05
s_{berg_PFT6}					0.5	± 0.01
$s_{berg\_PFT_7}$					2.9	± 0.03
s_{berg_PFT8}					0.48	± 0.01
$s_{berg\_PFT_9}$					0.69	± 0.01
$s_{berg\_PFT_{10}}$					1.36	±0.01
$s_{berg\_PFT_{11}}$					2.5	± 0.11
soil moisture						
$wSoil_{max(2)}$	752	± 0.02				
$s_{RD(1)}$			0.01	± 0.00		
$s_{RD(2)}$			0	± 0.00		
$s_{RD(3)}$			0.15	± 0.06		
$s_{RD(4)}$			0.15	± 0.07		
$wSoil_{max(RD_4)}$			145	± 0.08		
$wSoil_{max\_PFT_0}$					1.57	± 8.94
$wSoil_{max\_PFT_1}$					0.78	± 10.23
$wSoil_{max\_PFT_2}$					1.01	± 0.41
$wSoil_{max\_PFT_3}$					1.27	± 1.42
$wSoil_{max\_PFT_4}$					0.5	± 0.5
$wSoil_{max\_PFT_5}$					0.54	± 0.32
$wSoil_{max\_PFT6}$					0.85	± 2.53
$wSoil_{max\_PFT_7}$					01.01	± 0.57
$wSoil_{max\_PFT8}$					1.45	± 2.72
$wSoil_{max\_PFT_9}$					0.56	± 1.07
$wSoil_{max\_PFT_{10}}$					0.39	± 0.2
$wSoil_{max\_PFT_{11}}$					0.7	± 3.23

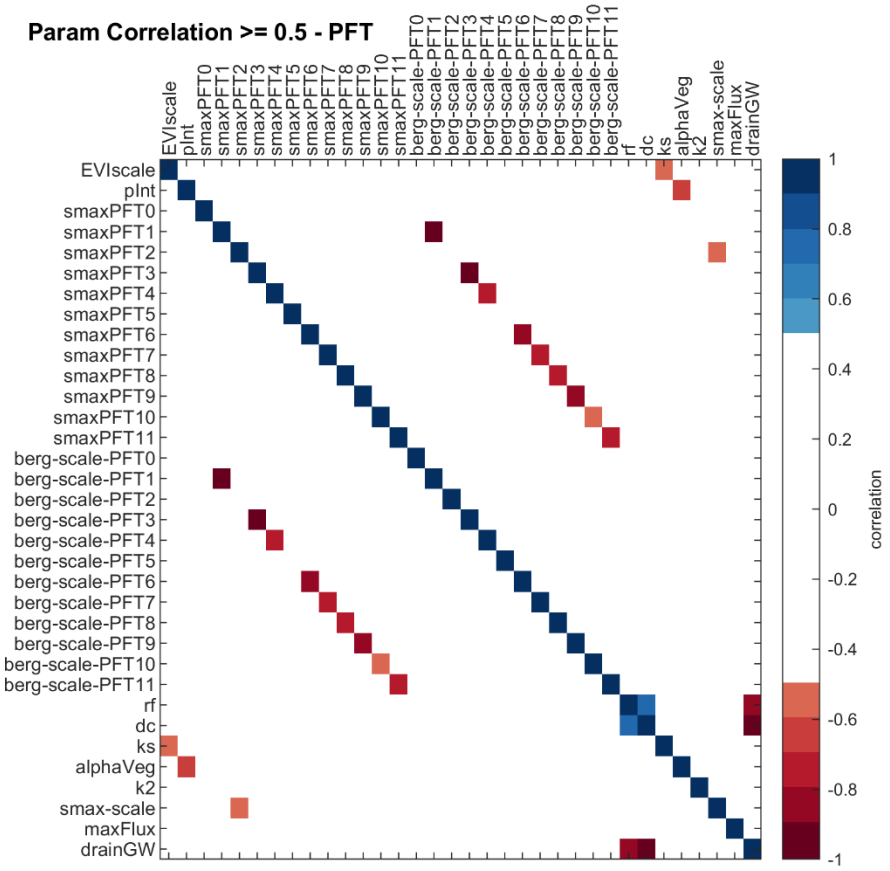

Figure 2: Correlation of calibrated parameters for the PFT experiment. Shown are only correlation coefficients |r|≥0.5.

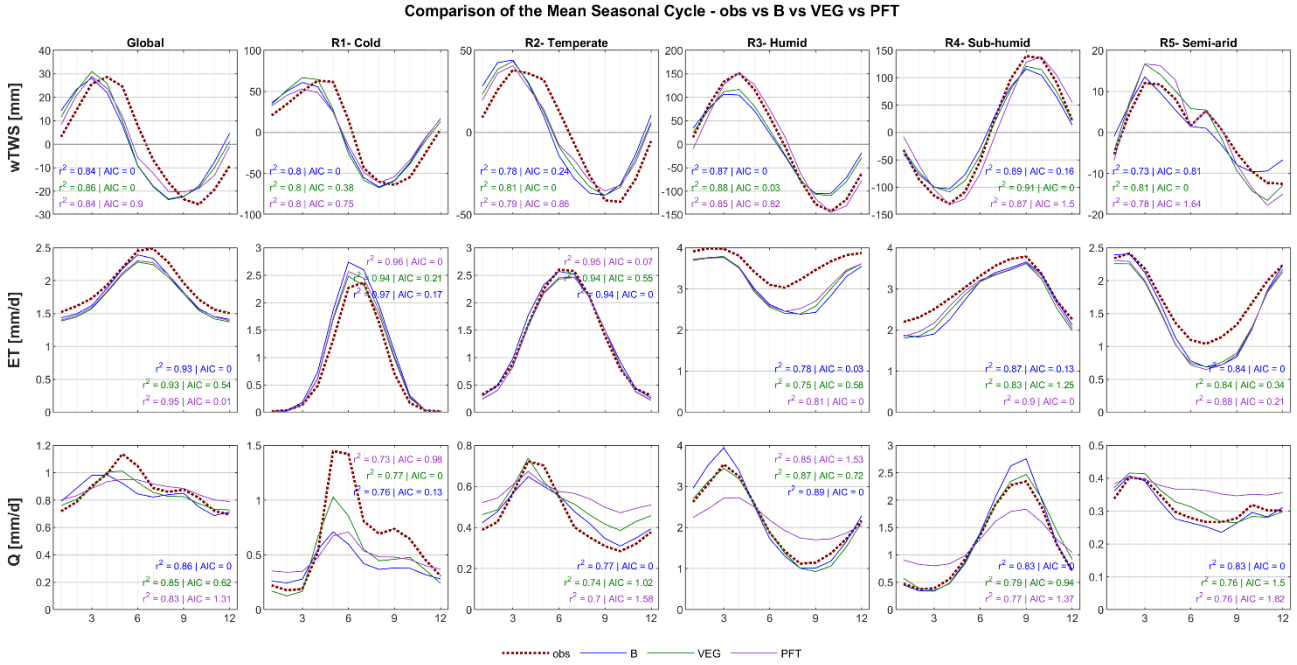

Figure 3: Global and regional mean seasonal cycles of total water storage (wTWS), evapotranspiration (ET) and runoff (Q) for the B, VEG and PFT experiments compared to the observational constraints by GRACE (wTWS), FLUXCOM (ET) and GRUN (Q). For each, the Pearson correlation (r^2) and Akaike information criterion (AIC) are calculated to compare model performance in terms of seasonal dynamics and of mean standard error in relation to the number of calibration parameters.

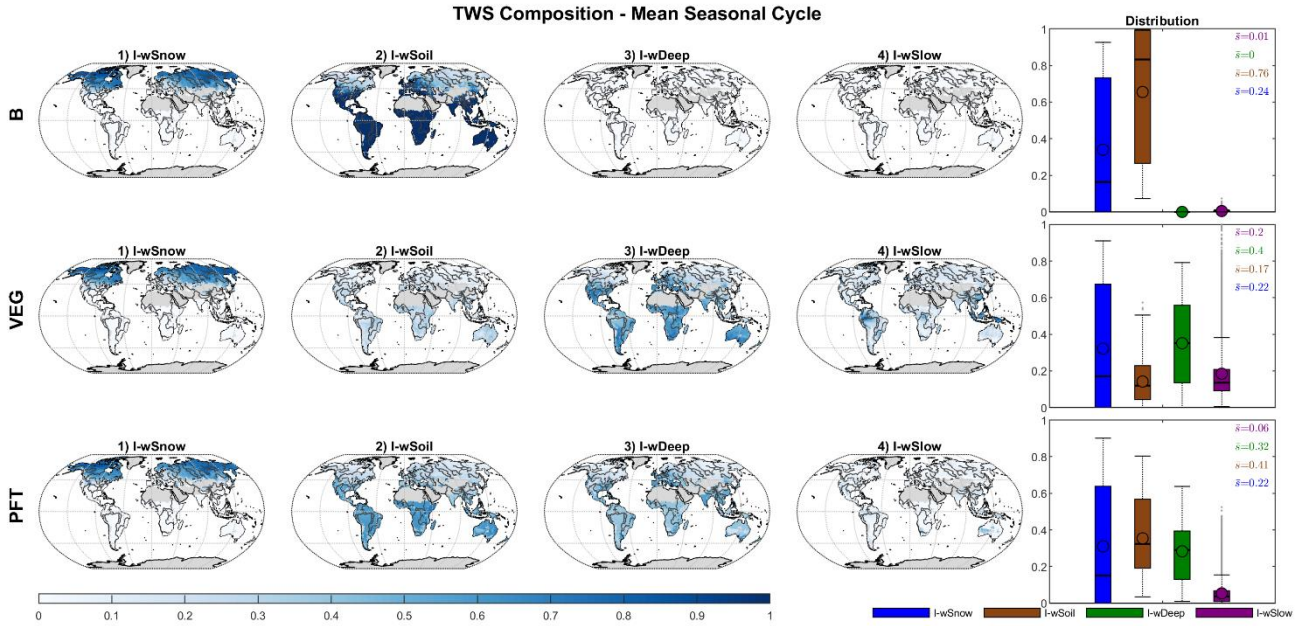

Figure 4: Global distribution of the Impact Index, I, for the contribution of simulated snow (wSnow), soil (wSoil), deep water storage (wDeep) and delayed water storage (wSlow) to the mean seasonal cycle of total water storage, for B, VEG and PFT.

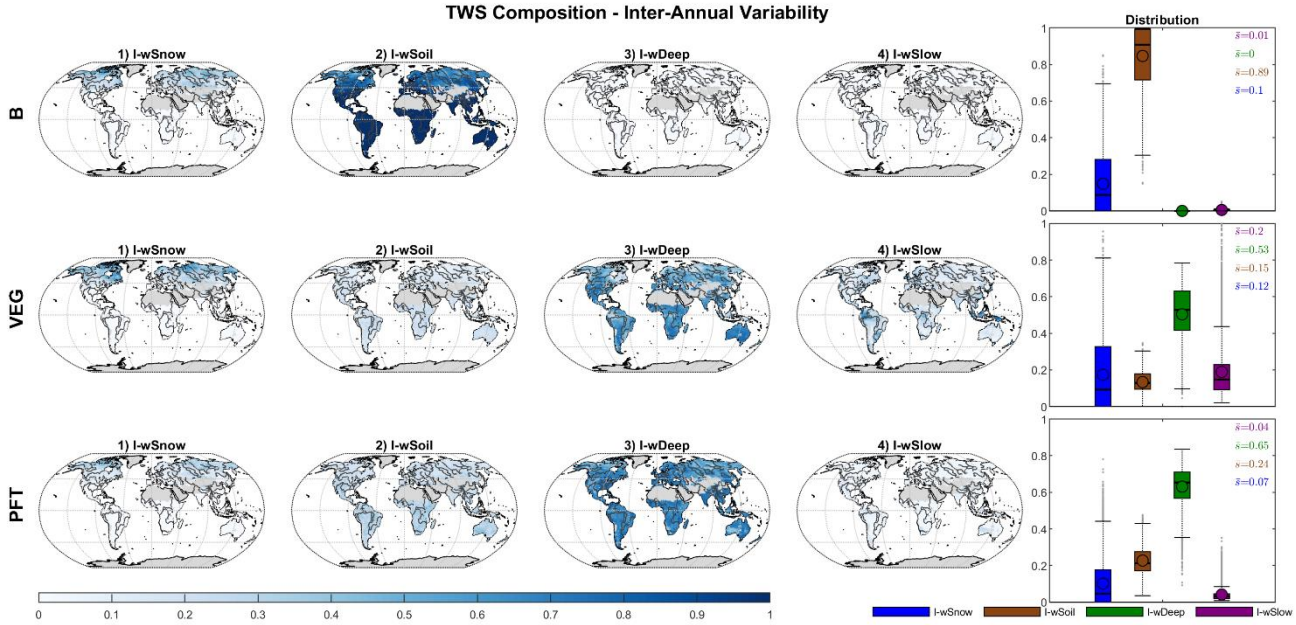

Figure 5: Global distribution of the Impact Index, I, for the contribution of simulated snow (wSnow), soil (wSoil), deep water storage (wDeep) and delayed water storage (wSlow) to the inter-annual variability of total water storage, for B, VEG and PFT.

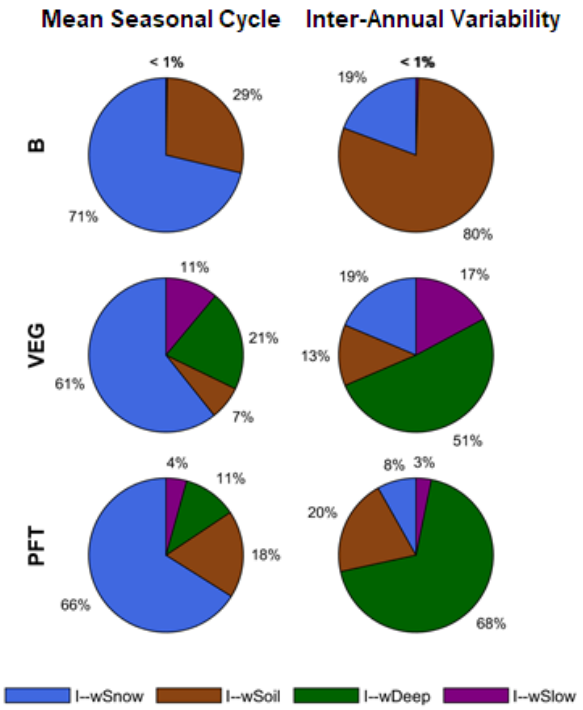

Figure 6: Impact Index, I, for the contribution of simulated snow (wSnow), soil (wSoil), deep water storage (wDeep) and delayed water storage (wSlow) to the global average mean seasonal cycle and inter-annual variability of total water storage, for B, VEG and PFT.

2) It is to note that some studies (see below and some literatures mentioned in the attachment) have dealt with the impact of dynamic vegetation on land surface processes, land-atmosphere interactions, etc. please help to discuss your novelty vs. what has been done.

- *Weiss, M., van den Hurk, B., Haarsma, R. et al. Impact of vegetation variability on potential predictability and skill of EC-Earth simulations. Clim Dyn 39, 2733–2746 (2012). https://doi.org/10.1007/s00382-012-1572-0*
- *BO Christoffersen, N Restrepo-Coupe, MA Arain …, Mechanisms of water supply and vegetation demand govern the seasonality and magnitude of evapotranspiration in Amazonia and Cerrado, Agricultural and Forest meteorology, 2014 https://doi.org/10.1016/j.agrformet.2014.02.008*
- *Weiss, M., Miller, P. A., van den Hurk, B. J. J. M., van Noije, T., Å□tefÄ□nescu, S., Haarsma, R., van Ulft, L. H., Hazeleger, W., Le Sager, P., Smith, B., & Schurgers, G. (2014). Contribution of Dynamic Vegetation Phenology to Decadal Climate Predictability, Journal of Climate, 27(22), 8563-8577.*

Also, this reviewer felt that the background/literature review part could be enhanced by citing some similar studies on using spatial information for model calibration, for example, those below.

- Ruiz-Pérez, G., Koch, J., Manfreda, S., Caylor, K., and Francés, F.: Calibration of a parsimonious distributed ecohydrological daily model in a data-scarce basin by exclusively using the spatio-temporal variation of NDVI, Hydrol. Earth Syst. Sci., 21, 6235–6251, https://doi.org/10.5194/hess-21-6235-2017, 2017.
- Su, Z., Zeng, Y., Romano, N., Manfreda, S., Francés, F., Ben Dor, E., ... Mannaerts, C. (2020). An integrative information aqueduct to close the gaps between satellite observation of water cycle and local sustainable management of water resources. Water, 12(5), 1-36. [1495]. https://doi.org/10.3390/w12051495

AC: We thank the Referee for a comprehensive suggestion on literature that helps to improve the background and clarify the motivation of our study. Following the suggestions, we will definitely include the references and adapt the introduction of the manuscript as follows:

> *[...The significance of interactions between vegetation and soil moisture are at the heart of ecohydrology (Rodriguez-Iturbe et al., 2001) and have become evident in many theoretical and experimental studies...]*
>
> Many studies analyzed effects of water availability on vegetation functioning (Porporato et al., 2004; Reyer et al., 2013; Wang et al., 2001; Yang et al., 2014), and the effect of changing vegetation cover on ecosystem water consumption (Du et al. 2021). While large-scale hydrologic models usually apply simplified and static vegetation characteristics (Quevedo et al. 2008, Weiss et al. 2012, Telteu et al. 2021), spatio-temporal variations of vegetation pattern are vital for good predictions of available water resources (Andersen et al. 2008). On ecosystem scale, Xu et al. 2016 showed the advantage of accounting for different plant hydraulic traits in an ecosystem model. And on a global scale, e.g., Weiss et al. 2012 showed the positive influence on modelled evaporation when replacing static vegetation characteristics by monthly LAI estimates in a climate model.
>
> However, how the representation of vegetation affects global water storages and in particular the partitioning of TWS in large-scale hydrological models has received little attention so far.
>
> *[...]*

Regarding the suggested literature on model calibration, we will include the suggested references and adapt the introduction in the revised manuscript as follows:

> *[...] This uncertainty of the available tools to interpret TWS variations is clearly a major obstacle for diagnosing and understanding global changes of the water cycle, which is increased by differing model structures and grown complexity of existing GHMs.*
>
> To improve model performance and reliability, hydrological models are traditionally calibrated against measured discharge time series at the outlet of catchments (Müller Schmied et al. 2021). However, discharge provides an integrated response of the catchment but not explicit evaluation of within-basin spatial heterogeneities. Therefore, the use of spatio-temporal data, e.g., from remote sensing, for model calibration has been suggested (Su et al, 2020). In fact, while a potential of using spatio-temporal data, of e.g., NDVI, could be shown at the catchment scale (Ruiz-Perez et al. 2017), many GHMs still have a limited usage of such data to calibrate model parameters. The most common approaches are still limited to tuning runoff-dependent model parameters against discharge observations of large catchments (Telteu et al. 2021). Some large-scale studies have shown clear improvements in model performance when a larger number of observational constraints are used to constrain the model parameters, especially when using terrestrial water storage variations from GRACE (e.g., Lo et al. 2010, Rakovec et al. 2016, Bai et al. 2018, Mostafaie et al. 2018, Trautmann, 2018). Among them, Trautmann et al. 2018
>
> *…contributed important insights in the drivers of TWS variations across spatial and temporal scales in northern high latitudes, [...]*

3) One major concern of this reviewer is that the use of various products for model calibration are not necessarily consistent. At least, the consistency issue should be checked and discussed before their use here. Sometimes, certain bias-correction might be needed to make various products consistent, before using them with the multi-criteria calibration approach. This reviewer also noticed that the author discussed a bit this in the discussion. Nevertheless, it is not fully clear how the inconsistency between different products will impact the output of the multi-criteria calibration.

AC: The Referee is right, inconsistency between the observational constraints is always an issue regardless of their usage in observation-based synthesis or as a data stream for model calibration. We would like to emphasize that this has been, at least partially, considered in the study. For example, we include the uncertainty of each data stream and focus on the most important and reliable patterns of each data stream in model calibration. For instance, we consider only the mean seasonal cycle of GRUN runoff due to its larger uncertainties reported on inter-annual scales (Ghiggi et al. 2019). Likewise, we focus on soil moisture dynamics instead of absolute values by using the Pearson' correlation coefficient as calibration criterion and further trim the considered soil moisture data.

Nevertheless, following the Referee's suggestion, we further assess possible inconsistencies between the different data products. Similar to the suggested study by Rodell et al. 2015 (see minor comment 9), we calculated the monthly water (im)balance, WB, from the observations for the period 01/2004-11/2010 (the time period in which none of the observation data has missing monthly values):

$$WB = P_{GPCP1DD} - ET_{FLUXCOM} - Q_{GRUN} - dS_{GRACE} \qquad \text{Eq. (AC1)}$$

with ideally WB = 0.

Fig. 7 shows the average monthly water imbalance scaled by each grid's average monthly precipitation $P_{GPCP1DD}$. While regionally large differences exist, the global mean and median are around 0. The global mean value of - 0.05 corresponds to a water balance residual of ~ 5% of precipitation - which is similar to the global residual of 4.3 % of precipitation reported in Rodell et al. 2014. Also temporally, the global average (Fig. 8) varies around 0, suggesting no major systematic inconsistency at the global scale, yet with a small imbalance with a tendency to negative values. This suggests that more water leaves the system than comes in when looking at the observational data. In comparison, there's obviously no imbalance when water balance is calculated with simulations from **B** and **VEG**, as they are based on water balance assumptions - which represents the major advantage of using models instead of observational based data from different sources.

We also calculated each variable in Eq. (AC1) by solving the water balance with the other observed components and compared the resulting water-balance-derived variable with the actual observed one. Differences between both indicate inconsistencies between a particular observed variable and the remaining observational variables. For ET, Q and wTWS, we additionally plot the modelled fluxes and storage changes from **B** and **VEG** to evaluate the effect of observational inconsistencies on model simulations (Fig. 8). The modelled fluxes are smoother and closer to the observations than the same estimate of the variable from the water balance. Therefore, we find that the model allows to potentially bridge the inconsistencies between the different data products. However, for dS, **B** and **VEG** show a time shift compared to the observed storage change, that isn't reflected in dS calculated from P, ET and Q observations. Accordingly, this underlines that the phase lag between observed and modelled

TWS variations is not caused by data inconsistencies, but rather related to the potential deficiencies in the model structure, as already discussed in the manuscript.

Fig. 9 compares the residuals of observed and simulated flux/storage change (*mod-obs*), and the ones of observed and water-balance derived variables (*WB-obs*). Large residuals between observed and water-balance derived variables point again to data inconsistencies of an observed variable with the remaining ones. When the residuals *WB-obs* and *mod-obs* in a region agree, it implies that the multi-criteria calibration approach prevents overfitting of the model(s) to an observed variable that is inconsistent with the remaining observed variables. Therefore, the model performance in these regions might be relatively poor against the inconsistent data stream, which is in fact a desirable behavior in the model calibration (e.g., ET in the *Semi-arid* region and dS in *Temperate* and *Humid* region).

When the residuals of *mod-obs* are considerably smaller than *WB-obs*, the model fits an observed variable well although it is inconsistent with the remaining observed variables (e.g., *Q* and dS in the *Semi-arid* region). Further, when the residuals of *mod-obs* are large but *WB-obs* doesn't indicate data inconsistencies, it points to issues related to model structure and parameter identifiability (e.g., Q in the *Cold* region, where the model(s) lacks the representation of permafrost, freeze/thaw dynamics and ice jam in rivers).

We will include these findings in the discussion of the revised manuscript and include the presented results in the supplement.

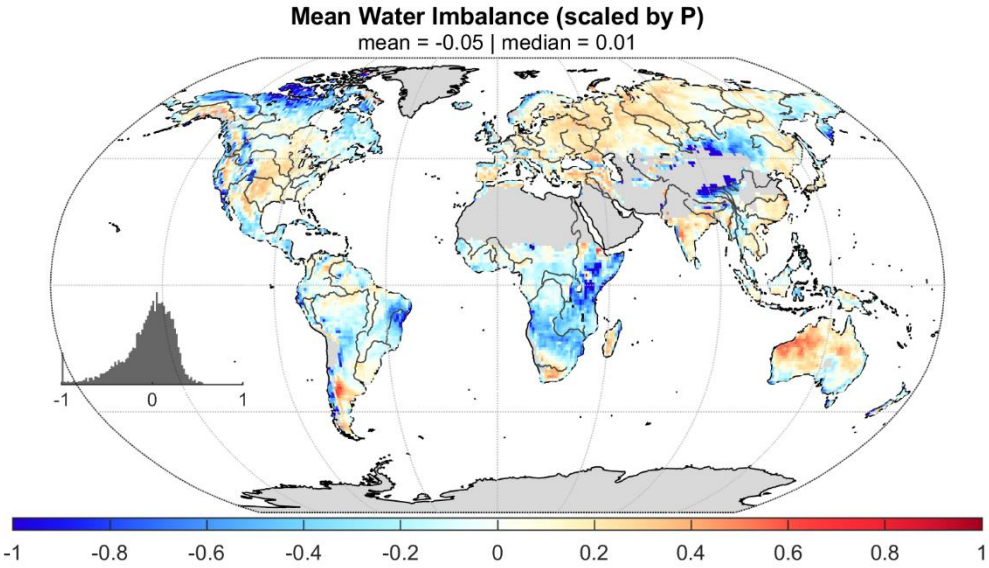

Figure 7: Mean water imbalance scaled by mean precipitation.

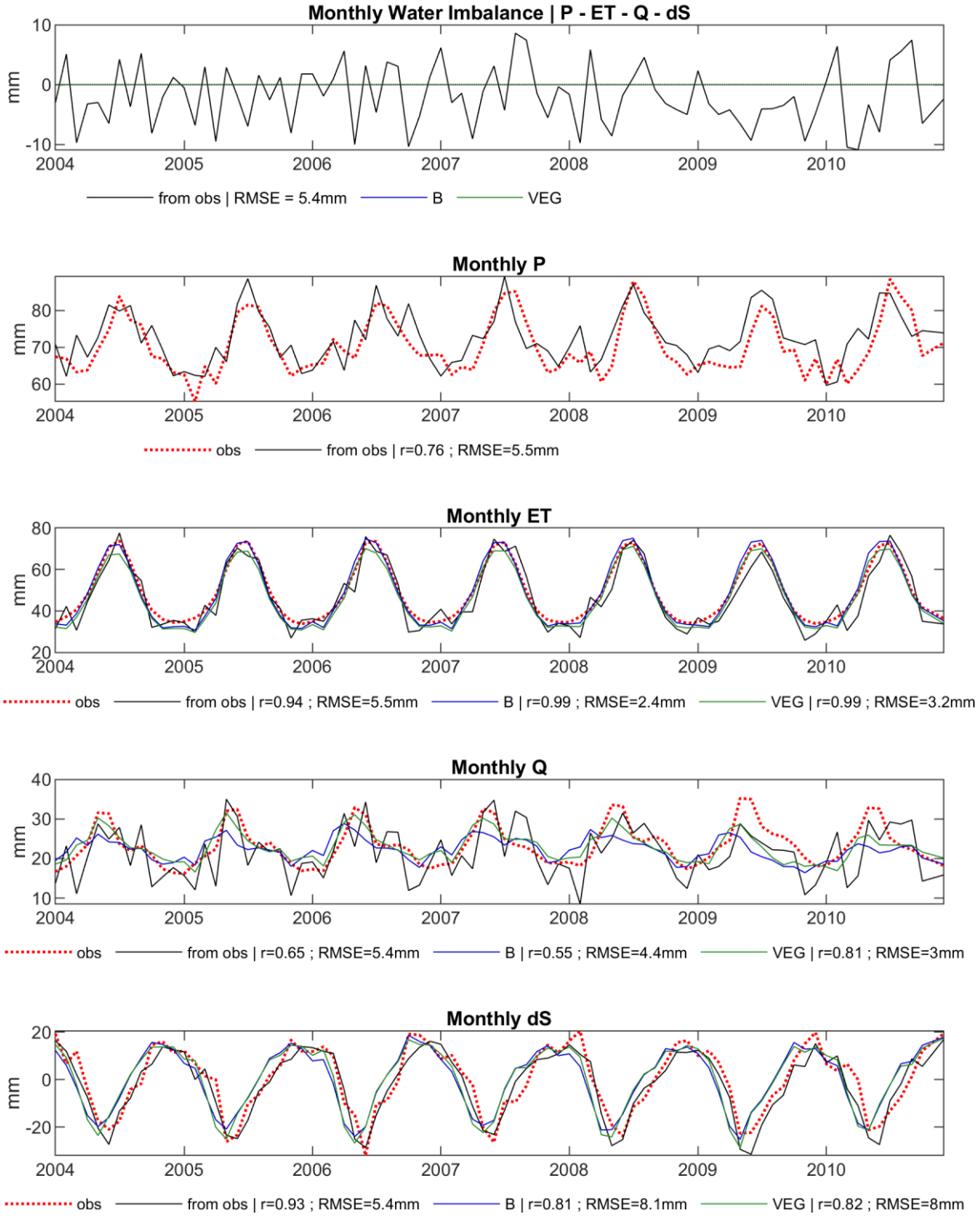

Figure 8: Global average time series of the water imbalance calculated from the observations (top row), and of water balance variables calculated from the other observations by resolving the water balance equations (from obs) vs the observed variable (obs) vs the simulated variable of the B and VEG simulations.

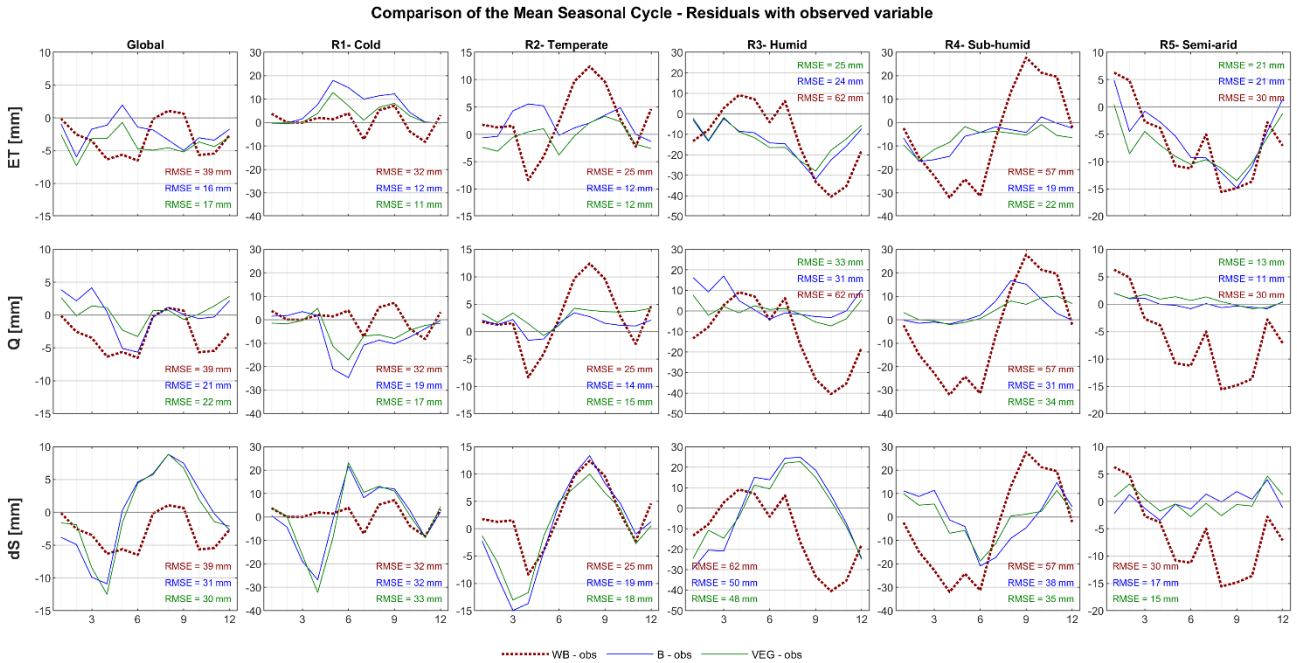

Figure 9: Global and regional mean seasonal cycle of the difference between observations and simulations from B and VEG, as well as difference between observed variable and the variable calculated via the water balance from the other observations, for ET, Q and dS.

Minor Comments

1) Line 46: It is worthwhile to mention that there are satellite-based root zone soil moisture products, using either data assimilation approach or analytical relationships between surface SM and root zone SM.

- *Reichle, R. H. , De Lannoy, G. J. M. , Liu, Q. , Ardizzone, J. V. , Colliander, A. , Conaty, A., Crow, W. , Jackson, T. J. , Jones, L. A. , Kimball, J. S. , Koster, R. D. , Mahanama, S. P. , Smith, E. B. , Berg, A. , Bircher, S. , Bosch, D. , Caldwell, T. G. , Cosh, M. , Holifield Collins, C. D. , Jensen, K. H. & 17 others, 2017, Assessment of the SMAP Level-4 Surface and Root-Zone Soil Moisture Product Using In Situ Measurements, Journal of hydrometeorology. 18, 10, p. 2621-2645*
- *Zhuang, R., Zeng, Y., Manfreda, S., & Su, Z. (2020). Quantifying long-term land surface and root zone soil moisture over Tibetan Plateau. Remote sensing, 12(3), 1-20. [509]. https://doi.org/10.3390/rs12030509*

AC: We agree with the Referee and will include the suggested references in the introduction as follows:

[...remote sensing-based estimates of soil moisture only capture depths up to 5 cm and do not necessarily reflect the moisture availability in the deeper soil column (Dorigo et al., 2015)]

While these observations can be extrapolated to derive estimates of root zone moisture, either by using statistical relationships (Zhuang et al. 2020) or by data assimilation into land surface models (Reichle et al. 2017, Martens et al. 2017), such products rely on the applied model.

[Therefore, GHMs are necessary to interpret TWS variations…]

2) Line 74: This reviewer think this is not under-studied. For example, see below refs.

- *Xu, X. T., Medvigy, D., Powers, J. S., Becknell, J. M. & Guan, K. Y. Diversity in plant hydraulic traits explains seasonal and inter-annual variations of vegetation dynamics in seasonally dry tropical forests. New Phytologist 212, 80-95, doi:10.1111/nph.14009 (2016)*
- *Du, L., Zeng, Y., Ma, L., Qiao, C., Wu, H., Su, Z. and Bao, G.: Effects of anthropogenic revegetation on the water and carbon cycles of a desert steppe ecosystem, Agric. For. Meteorol., 300, 108339, doi:10.1016/j.agrformet.2021.108339, 2021*

[... the inverse pathway of how vegetation properties influence dynamics of water pools and the partitioning of TWS in large scale models has received surprisingly little attention. …]

AC: The authors thank the Referee for highlighting these studies. We agree that the formulation of this sentence was not appropriate. We adjust the paragraph and will clarify that we refer to global studies as written in the author's response to major comment 2).

3) Line 83: why not add one more experiment to reflect the current/traditional approach in most of ESMs?

AC: This is a very good suggestion. We have performed an experiment that is much more comparable to the traditional approach. Please refer to the detailed response in major comment 1).

4) Line 96: This is very short section of 'method'.

AC: The referee is correct. The section numbering got mixed up. As Referee #3 also suggested, it should be *2. Methods*, *2.1. Overview* and then continue with *2.3 Model Description*. We will correct the section numbering accordingly.

5) Line 126: This reviewer believes this part of model description can be summarized with a paragraph with key characteristics, and then put the rest of detailed description to the appendix.

AC: We agree that the model description is quite long and detailed and summarizing it would improve the flow. However, we think that some of the equations, especially those that are 'changed' in the **VEG** experiment, are better suited in the main text, as they help to explain the model's behavior and to clarify the differences between (contributions of) water storages. We, therefore, will shorten the model description and only include the major equations in the revision. The revised model description section would read as follows:

2.3 Model Description

The conceptual hydrological model is forced by daily precipitation, air temperature and net radiation (Table 1). It includes a snow component (see Trautmann et al. (2018)), a 2-layer soil water storage (*wSoil*), a deep soil water storage (*wDeep*) and a delayed, slow water storage (*wSlow*). The schematic structure of the model is shown in Fig. 1 and calibration parameters are explained in Table 2.

Depending on air temperature (*Tair*), precipitation (*Precip*) is partitioned into snow fall (*Snow*), that accumulates in the snow storage (*wSnow*), and rainfall (*Rain*), that partly is retained in an interception

storage. Interception throughfall together with snow melt are distributed among soil through infiltration and infiltration excess depending on the ratio of actual soil moisture and maximum soil water capacity following Bergström 1995:

$$I_{exc} = I_{in} \cdot \left[\frac{\sum_{l=1}^{2} wSoil(l)}{\sum_{l=1}^{2} wSoil_{\max(l)}} \right]^{p_{berg}} \quad (1)$$

where, I_{exc} is the infiltration excess, I_{in} is the incoming water from throughfall and snow melt, *wSoil(l)* is the soil moisture and $wSoil_{max(l)}$ the maximum soil water capacity of each soil layer *l*, and p_{berg} is a global calibration parameter.

Part of the infiltration excess then replenishes a delayed water storage (*wSlow*), that acts as a linear reservoir and generates slow runoff (Q_{slow}). The remaining infiltration excess represents fast direct runoff (Q_{fast}). Q_{fast} and Q_{slow} together represent total runoff *Q*, that flows out of the system, i.e., grid cell.

Infiltrated water is distributed among 2 soil layers following a top-to-bottom approach, where the maximum capacity of the first soil layer is prescribed as 4 mm, in order to match the tentative depth of satellite soil moisture observations, while the storage capacity of the 2nd soil layer is a calibration parameter ($wSoil_{max(2)}$). The 2nd soil layer is connected with a deeper water storage (*wDeep*). The size of *wDeep* is defined as a multiple of $wSoil_{max(2)}$ by the calibrated scaling parameter s_{deep}. Depending on the moisture gradient between the two storages, water either percolates from the 2nd soil layer to the deeper soil, or it rises from the deeper storage into the 2nd soil layer, limited to a maximum flux rate. The deeper storage therefore acts as a storage buffer that linearly discharges further to the delayed water storage (*wSlow*). The *wSlow*, which also receives part of the infiltration excess, is thus representative of all delayed storage components.

Evapotranspiration (*ET*) is represented by a demand-supply approach that is driven by a potential ET demand following Priestley-Taylor, and is limited by the available soil moisture supply. The ET is partitioned into interception evaporation (E_{int}), bare soil evaporation from the first soil layer (E_{Soil}) and plant transpiration from the two soil layers (E_{Transp}). Interception and plant transpiration are only calculated for the vegetated fraction of each grid cell, while bare soil evaporation is limited to the non-vegetated fraction of each grid.

While water in *wSoil* is directly available for *ET*, *wDeep* is only indirectly accessible by capillary rise, and the water stored in *wSlow* is not plant-accessible. Total water storage is the sum of all water storages, including *wSnow*, *wSoil*, *wDeep* and *wSlow*. Although groundwater and surface water storages are not implemented explicitly, they are effectively included in *wDeep* and *wSlow*, especially after calibration of associated storage parameters against GRACE TWS.

6) Line 202: this could be another subsection, and not necessarily under the section of 'model description'.

AC: Thank you for the suggestion. We put *2.2.1 Including Vegetation characteristics* as a subsection of *2.2 Model Description*, because it describes the (**VEG**) model as well. Referee #3 made a similar suggestion. In order to

reduce the large numbers of sections and subsections, we will follow the suggestion and separate the sections *2.2 Model Description* and *2.3 Including Vegetation characteristics*.

7) Line 207: Do you know the below literature?

- *Ruiz-Pérez, G., Koch, J., Manfreda, S., Caylor, K., and Francés, F.: Calibration of a parsimonious distributed ecohydrological daily model in a data-scarce basin by exclusively using the spatio-temporal variation of NDVI, Hydrol. Earth Syst. Sci., 21, 6235–6251, https://doi.org/10.5194/hess-21-6235-2017, 2017.*

AC: This is a good point and we thank the Referee for suggesting this reference. We will include it in the introduction as also suggested in major comment 2) above.

8) Line 257: it is not clear how the p_berg is used to partition infiltration/runoff.

AC: p_{berg} is the runoff/infiltration coefficient that partitions incoming water, e.g., from throughfall and snow melt, into infiltration and infiltration excess (i.e., land surface runoff) based on the relative saturation of soil moisture, as shown in Eq. (1) of the manuscript. For a given maximum soil water capacity, p_{berg} = 1 means a linear relation between soil water saturation and the amount of incoming water that runs off: if the soil water pool is empty, most of the water infiltrates, whereas there is more infiltration excess when the soil is relatively saturated (see Fig. 10). Due to the exponential formulation of Eq. (1), p_{berg}<1 allocates a higher fraction of the incoming water to infiltration excess even if the soil water pool is nearly empty. On the contrary, p_{berg}>1 allows a large fraction of incoming water to infiltrate into the soil water pool when soil saturation is already high.

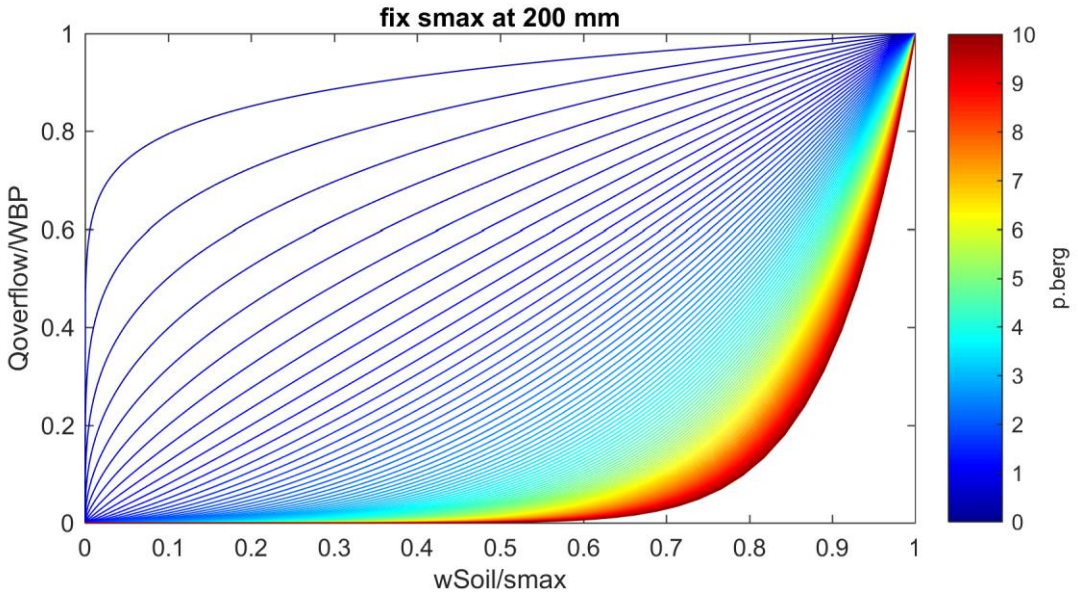

Figure 10: Relationship between relative soil saturation (wSoil/smax) and the fraction of infiltration excess (Qoverflow/WBP) as defined by different values of p_{berg}. Infiltration is then calculated as WBP-Qoverflow (WBP = Water Balance Pool, i.e. incoming water from rain fall and snow melt; Qoverflow = infiltration excess).

9) Line 273: It is not clear how consistent they are with each other, before using these observations as constraints. This reviewer think the consistency issue (see below refs.) among different products needs to be addressed before the use of them to constrain other models. At least, some discussions should be focused on this perspective. Also, please help to discuss how different products can affect your results, discussions, and conclusions.

- *Zeng Y., Z. Su, J.-C. Calvet, T. Manninen, E. Swinnen, J. Schulz, R. Roebeling, P. Poli, D. Tan, A.Riihelä, C.-M. Tanis, A.-N. Arslan, A. Obregon, A. Kaiser-Weiss, V. John, W. Timmermans, J.Timmermans, F. Kaspar, H. Gregow, A.-L. Barbu, D. Fairbairn, E. Gelati, C. Meurey, (2015) Analysis of current validation practices in Europe for space-based Climate Data Records of Essential Climate Variables, International Journal of Applied Earth Observations and Geoinformation, Vol 42, pp: 150-161, DOI: 0.1016/j.jag.2015.06.006*
- *Zeng Yijian, Zhongbo Su, Iakovos Barmpadimos, Adriaan Perrels, Paul Poli, K. Folkert Boersma, Anna Frey, Xiaogang Ma, Karianne de Bruin, Hasse Goosen, Viju John, Rob Roebeling, Joerg Schulz, Wim Timmermans, 2019, Towards a Traceable Climate Service: Assessment of Quality and Usability of Essential Climate Variables, Remote sensing, 11(10), 1-28. [1186]. https://doi.org/10.3390/rs11101186*

There are also a study below indicating how to evaluate water cycle products consistently:

- *Rodell, M., Beaudoing, H. K., L'Ecuyer, T. S., Olson, W. S., Famiglietti, J. S., Houser, P. R., Adler, R., Bosilovich, M. G., Clayson, C. A., Chambers, D., Clark, E., Fetzer, E. J., Gao, X., Gu, G., Hilburn, K., Huffman, G. J., Lettenmaier, D. P., Liu, W. T., Robertson, F. R., Schlosser, C. A., Sheffield, J. and Wood, E. F.: The observed state of the water cycle in the early twenty-first century, J. Clim., 28(21), 8289–8318, doi:10.1175/JCLI-D-14-00555.1, 2015.*

AC: We thank the Referee for raising a very important issue and suggesting suitable literature. As described in the response to major comment 3), we have considered the potential inconsistency issue, and provided an additional analysis based on the methodology from Rodell et al. 2015. In addition, in the revised manuscript, we will highlight the need to consider individual uncertainties and (processing) characteristics of each data set when interpreting the data by including the following paragraph:

> *[...] The parameters of each model variant are simultaneously optimized against multiple observational constraints, including monthly TWS anomalies from GRACE (Wiese et al. 2018), ESA CCI Soil Moisture (Dorigo et al., 2017), evapotranspiration estimates from FLUXCOM-RS ensemble (Jung et al., 2019) and gridded runoff from GRUN (Ghiggi et al., 2019) (Table 1) [..]*
>
> When using observational data sets from varying sources, it is essential to take into account the data's characteristics and uncertainties (Zeng et al. 2015, Zeng et al. 2020). Therefore, we calculate a cost term for each of the observational constraints, that considers the data's specific strengths and uncertainties, *[…]*

10) Line 311: There are several thresholds used to 'trim down' the study area for calibration. It would be nice to show all these percentages in one common map.

AC: We thank the Referee for this suggestion and agree that such a map would improve clarity on which grid cells were considered in the analysis and contributed to the results. Accordingly, we modified Figure 2 of the

manuscript by indicating cells that have not been included according to the different criteria explained in the figure caption (see Fig. 11).

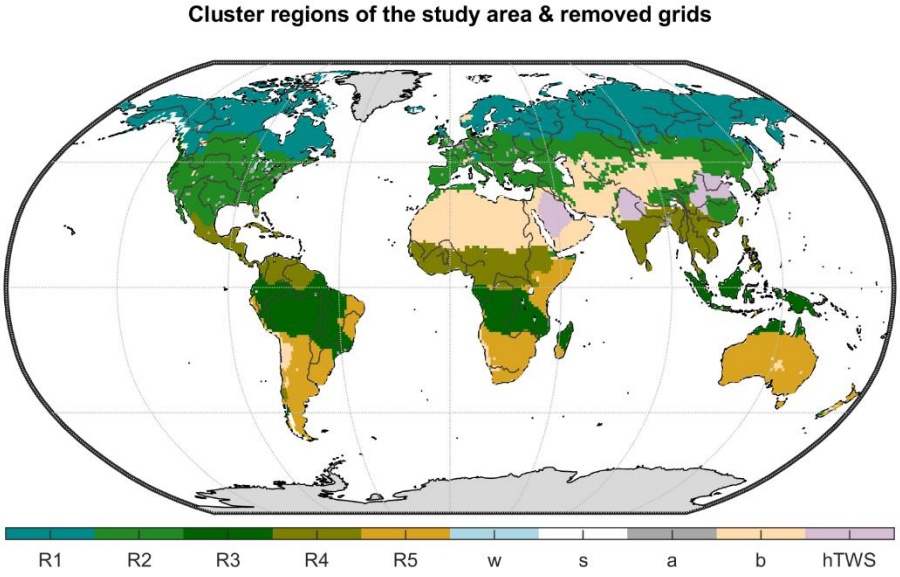

Figure 11: Hydroclimatic cluster regions of the study area (R1 - Cold, R2 - Temperate, R3 - Humid, R4 - Sub-humid, R5 - Semi-arid), and grid cells that have been excluded from this study (w = water fraction >50%; s = permanent snow and ice cover > 10%; a = artificial land cover fraction > 10%; b = bare land surface > 20%; hTWS = direct human impact on the trend in GRACE TWS).

11) Line 371: mode ?

AC: Here we refer to the modal value of the fraction of soil moisture that's available for evaporation, as reported in McColl et al. 2017. We thank the referee for pointing to this unclear expression. To improve the clarity, we will replace 'mode' by 'modal value' in the revised manuscript.

12) Line 381: Although it is understandable that the discussion linked to rooting depth is limited to the model structure, it is still worthwhile to discuss the combined control of precipitation and groundwater depth on rooting depth in various climate zones. See below ref:

- *Hydrologic regulation of plant rooting depth Ying Fan, Gonzalo Miguez-Macho, Esteban G. Jobbágy, Robert B. Jackson, Carlos Otero-Casal Proceedings of the National Academy of Sciences Oct 2017, 114 (40) 10572-10577; DOI: 10.1073/pnas.1712381114*

AC: We thank the Referee for this suggestion and will include the combined control on rooting depth in the introduction as follows:

> *[...] For example, vegetation promotes infiltration over surface runoff due to larger surface roughness, dampened precipitation intensities, more soil macro pores due to rooting and biological activity. In fact, such roles of vegetation in a global climate model were already envisioned and evaluated almost 4 decades ago (Rind, 1984).*

Besides, vegetation alters soil properties like soil texture and organic matter content. Such soil properties together with the plant's rooting depth control the size of the soil moisture reservoir that is available for ET, and how plants respond to drought stress conditions (Baldocchi et al., 2021;Yang et al., 2020). However, roots not only determine water supply for transpiration, but deep roots connect groundwater and provide access to the deeper moisture storages, influence the land-atmosphere interactions and thus have wider implications on the hydrological cycle. Rooting depth on the other hand is not only species-specific, but also determined by the precipitation infiltration and groundwater table depth depending on the topographic position, and presents a very large spatial heterogeneity not only across the globe, but locally as well (Fan et al. 2017).

The significance of interactions between vegetation and soil moisture are at the heart of ecohydrology (Rodriguez-Iturbe et al., 2001). [...]

13) Line 410: Fig.2 = Fig. 3?

AC: Thanks for pointing out the reference to the wrong figure. We will change accordingly.

14) Line 427: what do you mean trade-off here? Please clarify.

["… suggesting a trade-off between the two different observation data streams. …"]

AC: The calibrated model achieves either good regional performance regarding ET or regarding Q, but it cannot match both data streams equally well. This can be interpreted as a trade-off where a gain cannot be achieved in one without a corresponding loss on the other. On the one hand, this may indicate inconsistencies between the data streams and/or larger uncertainty in one of the data streams for a given region. On the other hand, it may indicate that relevant processes are missing in the model representation, and thus not allowing for both variables to improve at the same time. In either case, the trade-offs point to disagreements between observed and modelled fluxes that cannot be solved by model calibration alone.

We will rephrase the sentence clearly in the discussion in the context of inconsistency between the calibration data.

15) Line 527: Please explain why so? and provide a citation?

["… It should, however, be noted that the observational EVI data used in the VEG experiment do have an imprint (of the effects) of irrigated agriculture in terms…"]

AC: The EVI data is calculated from surface reflectance that are measured by remote sensing. By that, they represent a snapshot of the surface conditions at a given time, and show vegetation activity independent of whether it is enabled by natural or anthropogenic water supply. So, although no explicit information about irrigation has been provided, the increase in vegetation activity due to irrigation is measured by variations in EVI. We will modify the sentence in revision to clarify this point as follows:

[...]

It should, however, be noted that the observational EVI data used in the **VEG** experiment do have an imprint of the effects of irrigated agriculture, as the measured surface reflectance also include the increase of vegetation activity due to irrigation. The related better representation of ET may be associated

[... with an improved simulation of wTWS variations in such regions in the VEG experiments.]

16) Line 537: what do you mean here? please clarify and make it explicitly.

["…the bias regarding either ET or Q, may relate to shortcomings in the precipitation forcing that doesn't provide sufficient input to support both outgoing water fluxes …"]

AC: This sentence also relates to comment 15). Here we additionally suggest that the bias regarding either one of the outgoing fluxes relates to the model's inability to allocate water to both water fluxes while maintaining the water balance. We attribute this to potential bias in precipitation which would not provide sufficient water. It is well known that global precipitation datasets are potentially affected by underestimation of precipitation (Trenberth et al. 2007, Contractor et al. 2020). Those limitations relate to the satellite measurements (sensor sensitivity to different precipitation types, retrieval methods, discontinuous nature of observations), gauge measurements (gauge network density, instrument sensitivity, local influences, wind/wetting/evaporation errors) and, if combined, from the combination method (Fekete et al. 2004). In this study, we use GPCP 1DD precipitation data that combines satellite infrared and microwave measurements that were scaled to ensure consistency with monthly gauge-based datasets. Validation of GPCP 1DD showed an underestimation of precipitation in complex terrain and regionally during spring and autumn, while precipitation in winter time tends to be overestimated (Huffman et al. 2001). While we account for the latter by reducing snowfall (via a scaling parameter that was calibrated in Trautmann et al. 2018), we don't consider potential underestimation.

Additionally, while monthly estimates are robust, daily precipitation values rely on assumptions in the temporal distribution of rainfall at sub-monthly time scales (Huffman et al. 2001, Herold et al. 2016), which influences simulated daily fluxes and feedbacks within the model that might lead to effects on the monthly time scale.

We will include the description in the revised manuscript accordingly.

17) Line 547: there are some latest studies on this perspective:

- *Yu, L., Zeng, Y., & Su, Z. (2020). Understanding the mass, momentum, and energy transfer in the frozen soil with three levels of model complexities. Hydrology and Earth System Sciences, 24(10), 4813-4830. https://doi.org/10.5194/hess-24-4813-2020*
- *Yu, L., Fatichi, S., Zeng, Y., and Su, Z.: The role of vadose zone physics in the ecohydrological response of a Tibetan meadow to freeze–thaw cycles, The Cryosphere, 14, 4653–4673, https://doi.org/10.5194/tc-14-4653-2020, 2020*

[... The remaining deficiencies in model performance, especially in the Cold region, indicate missing processes in the simple model structure. Such processes include freeze/thaw dynamics, permafrost and ice jam in river channels that would increase surface water storage and allow high spring flood. …]

AC: Thank you for suggesting these studies on the role of vadose zone physics representation on simulated ecohydrological responses. We will include the references in the revised manuscript.

18) Line 678: remove last sentence

> *[… Besides, this study motivates further multi-model experiments to understand the need and potential of existing and novel observational constraints to increase the identifiability not only regarding model parameters, but also of model structure. …]*

AC: The sentence will be removed.

19) Line 687: Why not make it open on GitHub?

AC: We agree with the Referee's suggestion and are currently archiving and sorting the code and data used for this analysis. In the revised manuscript, we will include a public link to access the code and data.

Author's response to Referee #2

1) **Line 148: Eq.2 should be Eq.3.**

   AC: Yes, the Referee is right. We will reference the correct equation in the revised manuscript.

2) **Line 230: typo, 'Therefor'.**

   AC: Thanks for finding the typo. We will correct it.

3) **Line 410: Fig.2 should be Fig.3.**

   AC: Once again, thank you. We will point to the correct figure in the revision.

4) **Figure 4: please add the label of y axis (numbers of grids?).**

   AC: Yes, the y-axis of the histograms in Fig. 4 shows the number of grid cells. We will add the label for the y-axis to make it clear.

5) **Line 488: please check the number '69%', I noticed that this number is 61% in Figure 8.**

   AC: Thanks for pointing this out. We will change accordingly in the revised manuscript, and ensure that such mistakes are not repeated.

Author's response to Referee #3

Major Comments

1) **This paper mainly focus on investigating the impact of "space and time varying vegetation parameters" on defining infiltration, root water uptake and transpiration processes and decomposing the TWS. It sounds strange to use the words such as "importance of vegetation", "including vegetation", "including vegetation characteristics", "including vegetation data" and "contribution of vegetation". These words cannot reflect well the research objective, and I think the impact of vegetation/vegetation characteristics/ vegetation data is also embedded within the baseline experiment using the static and globally uniform parameter values.**

AC: We thank the Referee for this constructive criticism and agree that the baseline experiment implicitly accounts for the role of vegetation in its globally uniform parameters. We clarified the statements mentioned by the Referee by rephrasing into "including vegetation explicitly", "including varying vegetation characteristics" or "importance of variation of vegetation properties" at the respective occasions. We further clarify this aspect by introducing the following sentences in *2.1 Overview*:

> *[... In the VEG experiment, we describe vegetation related parameters as the linear product of a calibrated scalar and spatio-temporal varying vegetation variables. By calibrating the scalar, we include the continuous pattern from the data, but weight it to best fit with observational constraints.]*
>
> Even though the optimized parameters of the baseline experiment implicitly account for the effect of vegetation, its parameters are global constants and do not vary spatially. In the **VEG** experiment vegetation related parameters vary explicitly spatially and partly temporally.
>
> *[...]*

2) **The authors calibrate and validate the model performance almost for the same period (01/2002-12/2014 vs. 03/2000-12/2014). From the view of traditional calibration and validation procedure, it will be better to use part of the observations (e.g. period of 2002-2008) to calibrate the model, and then using the remaining observations (e.g. period of 2008-2014) to validate the model.**

AC: This is a good point. Due to the rather short time period of observational constraints (01/2002-12/2014, which are 144 monthly values), we decided to not further reduce the temporal information. Instead of splitting the time period, we divided the calibration and validation data spatially. This is similar to the proxy basin test in traditional catchment hydrology, when a model is calibrated for one basin and evaluated for another. We will add a sentence to the revised manuscript to clarify this.

3) **For the regional analysis of model performance, the authors derive 5 hydroclimatic regions by performing a cluster analysis, but it sounds strange to treat almost the whole China and Europe as the same group, i.e., the moderate mid latitudes (Temperate). This is contrast to the common sense. The authors are thus suggested to use a better classification such as the Köppen–Geiger climate classification.**

AC: Our regional classification was based on clusters of seasonal dynamics of the observation data used in this study (ET, Q, and TWS). It additionally includes the latitude along which the main gradient of the regional climate exists. Note that the classification was agnostic to geographical regions or boundaries, and the regions showing similar hydro-climatic features were grouped into a cluster. We are aware that it is common to use the Koeppen-Geiger classification for regionalization. However, one of the major shortcomings of Koeppen-Geiger classes is the inclusion of grid cells from Southern and Northern latitudes into one class. These regions with opposing seasonal cycles potentially distort the seasonality of hydrological variables at the regional scale. We will therefore keep the regionalization based on hydroclimatic clusters as the basis of our analysis and results.

Nevertheless, following the Referee's suggestion, we performed a regional analysis for Koeppen-Geiger climate zones instead of hydroclimatic clusters. To do so, we aggregated Koeppen-Geiger subgroups considering the main climate group and distinguishing between humid and semi-arid conditions. The resulting zones are shown in Fig. 1. Fig. 2 evaluates model performance for the Koeppen-Geiger regions and Fig. 3 shows the composition of seasonal TWS variations therein. The results presented below will be included in the supplement of the revised manuscript.

Note that most parts of the *Polar* and *Boreal* Koeppen-Geiger (KG) zone are included in the *Cold* region (R1) of the hydroclimatic cluster classification. We find that the regional averages are very similar for both classification schemes in terms of model performance and composition of seasonal TWS variations.

The Northern Hemisphere *Temp* and *Boreal-sa* KG zones are both included in the *Temperate* hydroclimatic region (R2). *Temp* KG and the *Temperate* region (R2) agree well regarding model performance and seasonal cycles, although we see a slightly better performance for the *Temp* KG regarding wTWS and Q. In the *Boreal-sa* KG, **B** and **VEG** don't reproduce the spring peak of Q and precede the observed wTWS significantly, decreasing model performance slightly when combining the *Temp* and *Boreal-sa* KG zones in one hydroclimatic region. Therefore, it would make sense to further distinguish the *Temperate* hydroclimatic cluster region. However, *Boreal-sa* KG spans Northern China, where poorer model performance is also evident from the performance maps in Fig. 4 of the manuscript.

However, as mentioned in the manuscript, the advantage of the hydroclimatic cluster regionalization becomes obvious when interpreting results of the *Arid* and *Temp-sa* KG zones. This is because these climate zones are distributed across the Southern and Northern Hemisphere, causing 2 peaks in the regional seasonal cycles for wTWS, ET and Q, due to opposing seasonal dynamics. The *Arid* KG zone includes the *Semi-arid* cluster regions (R5) in the Southern Hemisphere, as well as parts of the *Temperate* region (R2) (mainly in North America). The *Temp-sa* KG zone covers a rather small fraction of the study area, that is spread over the *Temperate* region (R2) in the Northern Hemisphere and the *Semi-arid* region (R5) of the Southern Hemisphere.

Likewise, the effect of opposing seasonal cycles is apparent in the *Tropic* KG zone, although less apparent due to the proximity to the equator where the climate is more homogeneous and seasonality is low. The *Tropic* KG corresponds to the *Humid* cluster region (R3) on the Southern Hemisphere, and parts of the *Sub-humid* region (R4) on the Northern Hemisphere. Compared to the hydroclimatic cluster regions, the *Tropic* KG has less seasonal variation (a smaller amplitude) of wTWS, ET and Q, due to its larger area North and South of the equator. Both, **B** and **VEG** underestimate the ongoing depletion of wTWS from September to December in *Tropic* KG, which is likely related to the opposing seasonal cycles of wTWS in the *Humid* (R3) and the *Sub-humid* (R4) cluster regions. In the *Tropic* KG, Q peaks in March (as in *Humid* (R3)) and has a second, smaller peak in September (when Q

peaks in the *Sub-humid* region (R4)). However, model performance is very similar for *Tropic* KG and the *Humid* and *Sub-humid* cluster regions.

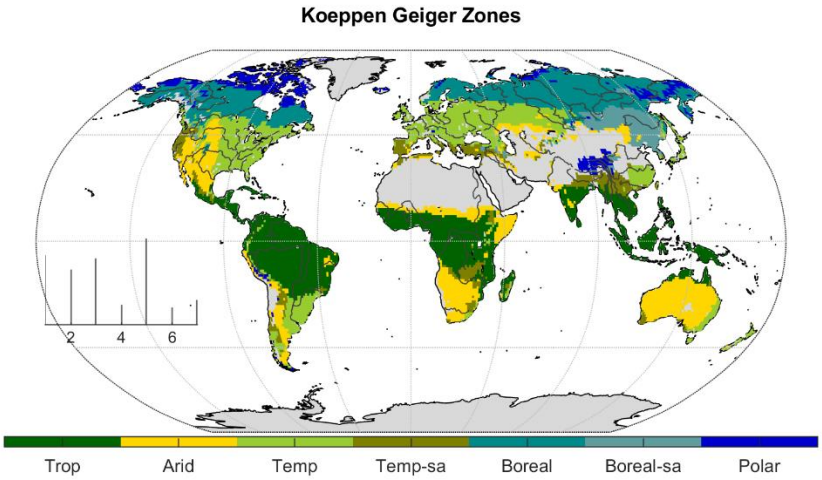

Figure 1: Regions based on Koeppen-Geiger climate zones (Trop = Af, Am, As, Aw; Arid = BSh, BSk, BWh, BWk; Temp = Cfa, Cfb, Cfc, Dfa, Dfb; Temp-sa = Csa, Csb, Csc, Cwa, Cwb, Cwc; Boreal = Dfc, Dfd; Boreal-sa = Dsa, Dsb, Dsc, Dwa, Dwb, Dwc, Dwd; Polar = EF, ET).

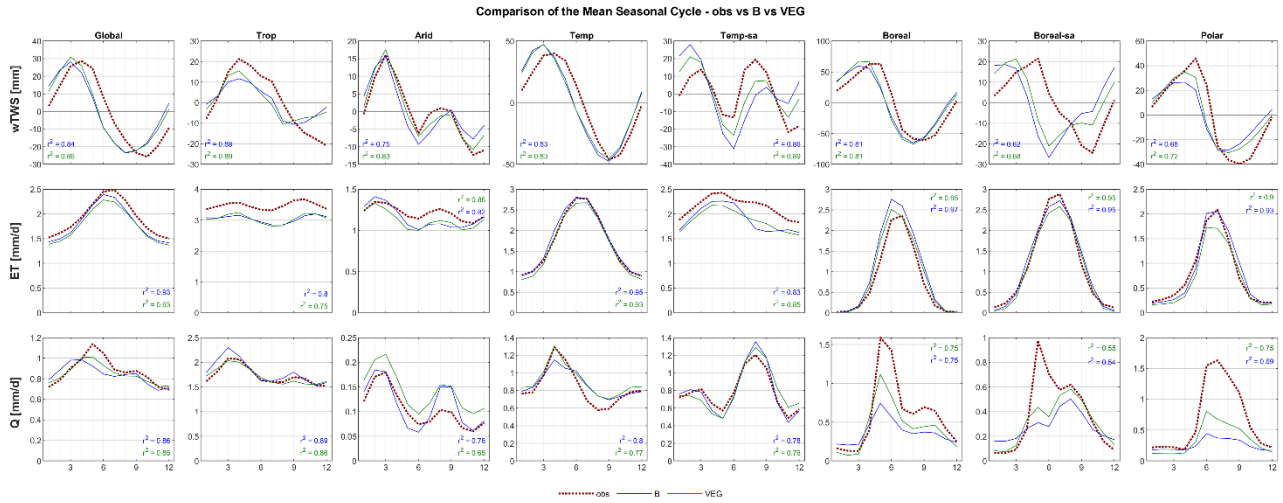

Figure 2: Global and regional mean seasonal cycles of total water storage (wTWS), evapotranspiration (ET) and runoff (Q) for the B and VEG experiments compared to the observational constraints by GRACE (wTWS), FLUXCOM (ET) and GRUN (Q).

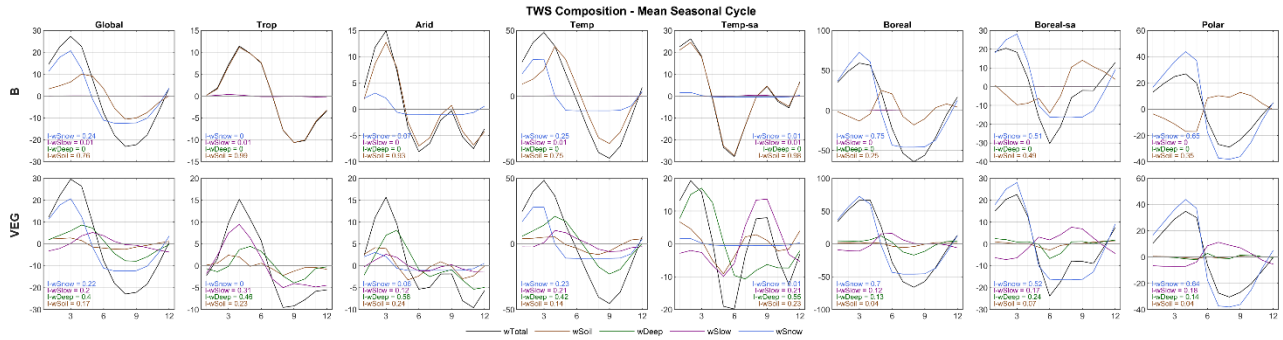

Figure 3: Global and regional mean seasonal cycles of simulated total water storage and its components for B and VEG, including the regional Impact Index *I* for each storage.

Minor Comments

1) **The title of Section 2 is wrong and the structure of this section can be improved. E.g., "1.1 Methods" should be "2 Methods", "1.2 Overview" should be "2.1 Overview". In addition, it's suggested to modifying "2.2.1" to "2.3".**

AC: We thank the Referee for pointing out the mixed-up section numbering. The suggestion also aligns with Referee #1's minor comments 4) and 6), and we will follow the suggestions in the revised manuscript.

2) **It's suggested to modifying all the tables to the standard table format.**

AC: Thanks for the suggestion! We included grey shaded rows as 'subsections' of the table to improve readability. However, we will adapt the layout of the tables and figures when and if requested by HESS.

3) **Line 410: "Fig. 2" should be "Fig. 3"**

AC: Thanks for pointing out the reference to the wrong figure. We will change accordingly.

4) **The number in Fig.3 is difficult to read**

AC: We will increase the font sizes of the numbers in Fig. 3.