Response to Reviewer #1

General comment: Globally, the paper is well written and structured. It is worth for publication, but minor revisions should be considered. For example, there are some results that are not shown in the paper (you mention these), but for the readers they do not know what it is. It could be better if these results can be shown more specifically.

We would like to thank Reviewer #1 for the positive feedback and constructive comments. We provided our responses (R, in blue) to each of the reviewer's comments (C, in black) below:

C1: 144 – 145: "temporal disaggregation is applied to CHIRP precipitation …". We know that precipitation is not a continuous variable. Disaggregation can lead to some unexpected errors. So why not using the satellite precipitation data that are sub-daily scale?

R1: We agree with Reviewer #1 that using sub-daily data should be ideal in our case. However, by the time we performed our analysis, only 0.05° global precipitation data (from CHIRPS) were available to us. As such, we used CHIRPS data and adopted the same temporal disaggregation as the previous research (e.g., McNally et al., 2017). Despite a possible additional precipitation uncertainty, using the disaggregated CHIRPS data in our development leads to an improved result. The 0.05° sub-daily global precipitation data can be used in the model simulation when they are available.

C2: 166 - 167: The temporal mean value of model simulated TWS is used to convert ΔTWS into absolute TWS. This could be accepted, but you use the period of 2003 to 2012. Why not use the whole simulation period? Because your simulation period is from 1981 to 2012.

R2: Reviewer #1 is correct that the model simulation (open-loop) is from 1981 to 2012. However, GRACE data are only assimilated between 2003 and 2012 due to their availability. The long-term mean value of 2003-2012 is used associated with the GRACE DA period. We reported the GRACE DA period in **lines 404 – 406** of the revised manuscript:

GRACE observations are assimilated into the CABLE 0.5° and CABLE 0.05° models (called GRACE DA 0.5° and GRACE DA 0.05°, respectively) between January 2003 and December 2012 (due to the availability of meteorological forcing and GRACE data).

To clarify this further, we also add a description of the GRACE DA period in Sect. 2.3, **lines 173 - 174**:

In this study, GRACE data are assimilated into CABLE between January 2003 and December 2012 (due to GRACE data availability).

C3: 301: What is the point of the phase estimates? Could you explain it more specifically?

R3: In this section, the phase estimate is used to demonstrate the spatial details of our model simulation. The phase shows more details than the amplitude, and we believe that showing both provides readers extra information on the improved spatial detail. The phase exhibits the timing when TWS reaches the maximum value (with respect to the beginning of the year). It illustrates the wet period of the year. For clarity, we add the description of the phase estimate in Fig. 5's caption, **lines 848 – 850**:

Figure 5: Annual amplitude (top) and phase (bottom) of the TWS estimates computed from CABLE 0.5° (a, c) and CABLE 0.05° (b, d). The insets in (a) and (b) show details in southeast Australia. The phase exhibits the timing when TWS reaches the maximum value (with respect to the beginning of the year). The unit of the phase is a calendar month, e.g., January (J), December (D).

C4: 307 – 308: For CABLE 0.5, the correlation length (CL) of August is the smallest. The CL of June is higher than that of seven months. For CABLE 0.05, the CL of May is the smallest but is not that small. Besides, the CLs of June and July is approximately equal to that of January and February. How to explain these?

R4: We thank Reviewer #1 for addressing this. Please note that the correlation length is computed across Australia, where the timing of wet/dry periods of TWS in different regions may also affect the correlation length estimates. Our analysis mainly relies on TWS variations in Northern and Southern regions, which are the most significant. Different spatial features of CABLE 0.5° and CABLE 0.05° model parameters may also play a role in the TWS spatial distribution. However, the overall temporal can be described as follows:

Wet and dry conditions lead to smoother (i.e., more uniform) spatial features, resulting in a larger correlation length. In wet seasons, an aquifer is slowly recharged and filled after several weeks (or months) of rainfall. This explains why the approximate peaks are observed around Feb-Mar and Sep-Oct. At the beginning of the wet season, scattered rainfall in part of the continent likely causes a gradient between dry/wet areas, resulting in a smaller correlation length (i.e., the spatial distribution is not as smooth as the very wet/dry case). The same is also observed at the end of wet or beginning of the dry season, e.g., May. Both Northern and Southern Australia are relatively dry in June, resulting in a larger correlation length.

For clarity, we modify our explanation as follows (please see **lines 327 – 333**):

…Larger correlation lengths are found during the rainy seasons (Jan – Apr in the North and Aug – Nov in the South) and during the dry season (e.g., Jun). Soil and aquifer storage increase during the wet seasons, leading to more uniform (and smoother) spatial moisture features. Similar uniformity can also be observed during the dry season. At the beginning of the wet season, scattered rainfall in part of the continent likely causes a gradient between dry/wet areas, resulting in smaller correlation lengths. It is noteworthy that our analysis only explains the overall temporal pattern of continental correlation lengths. The temporal pattern may also be affected by the local TWS wet/dry features or by the spatial distribution of model parameters.

C5: 385 – 386: It is not surprised to obtain this result because you add the information of GRACE into the model. More information will lead to more similarities.

R5: Reviewer #1 is perfectly correct.

C6: 399: “You state that the GWS is the primary driver of the TWS trend in Line 321. However, the result show that the SMS is a major contributor of TWS. Could you explain it?

R6: Reviewer #1 is correct. The trend of TWS is mainly governed by GWS (line 321). However, the annual amplitude of TWS is dominated by SMS. To clarify this, we modify the statement in Sect. 4.2 to:

… The contribution is calculated as a percent of the annual amplitude of TWS fluctuations. In CABLE 0.05° (Fig. 13a), the SMS is a major contributor to more than 90% of the TWS variation (i.e., annual amplitude). …

The revised statement can be found in **lines 420 – 422**.

C7: 399 – 400: The GRACE mostly shows or indicates the changes of groundwater, so it is apparent to have this result.

R7: We thank Reviewer #1 for the suggestion. The results are shown in Fig. S1 (below figure):

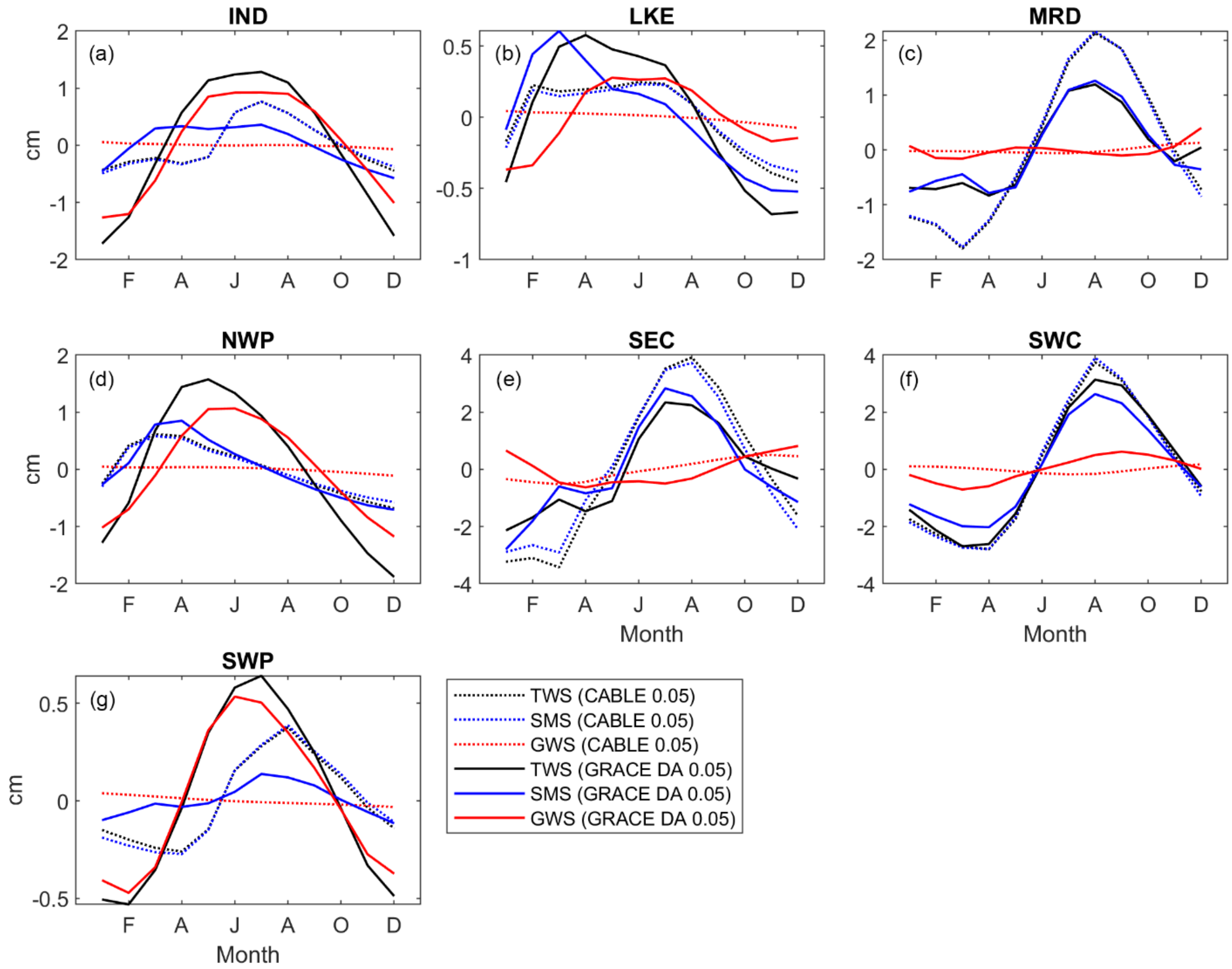


Fig. S1: Monthly basin averaged TWS, SMS, and GWS variations from CABLE 0.05° and GRACE DA 0.05°.

However, we found that Fig. S1 might be redundant to Fig. 13 (TWS contribution in the main text). As readers can also access this report, we decide to show only Fig. 13 in the paper to avoid redundancy.

C8: 428: A positive value means the former (vertical) is better than the latter (horizontal), right? So the negative value (-0.1) implies that CABLE 0.05 (OL) has a lower correlation than the GRACE DA 0.5 (DA), if I understand it correctly.

R8: We highly appreciate Review #1 for pointing this out. We accidentally switched the order between GRACE DA 0.5° and CABLE 0.05° here. We will correct the error as follows:

… we find that GRACE DA 0.5° (DA) shows a higher correlation value than CABLE 0.05° (OL) by 0.1 (see Fig. 14d3). This indicates that improving model state estimates via DA is more effective than improving model parameters via increased resolution.

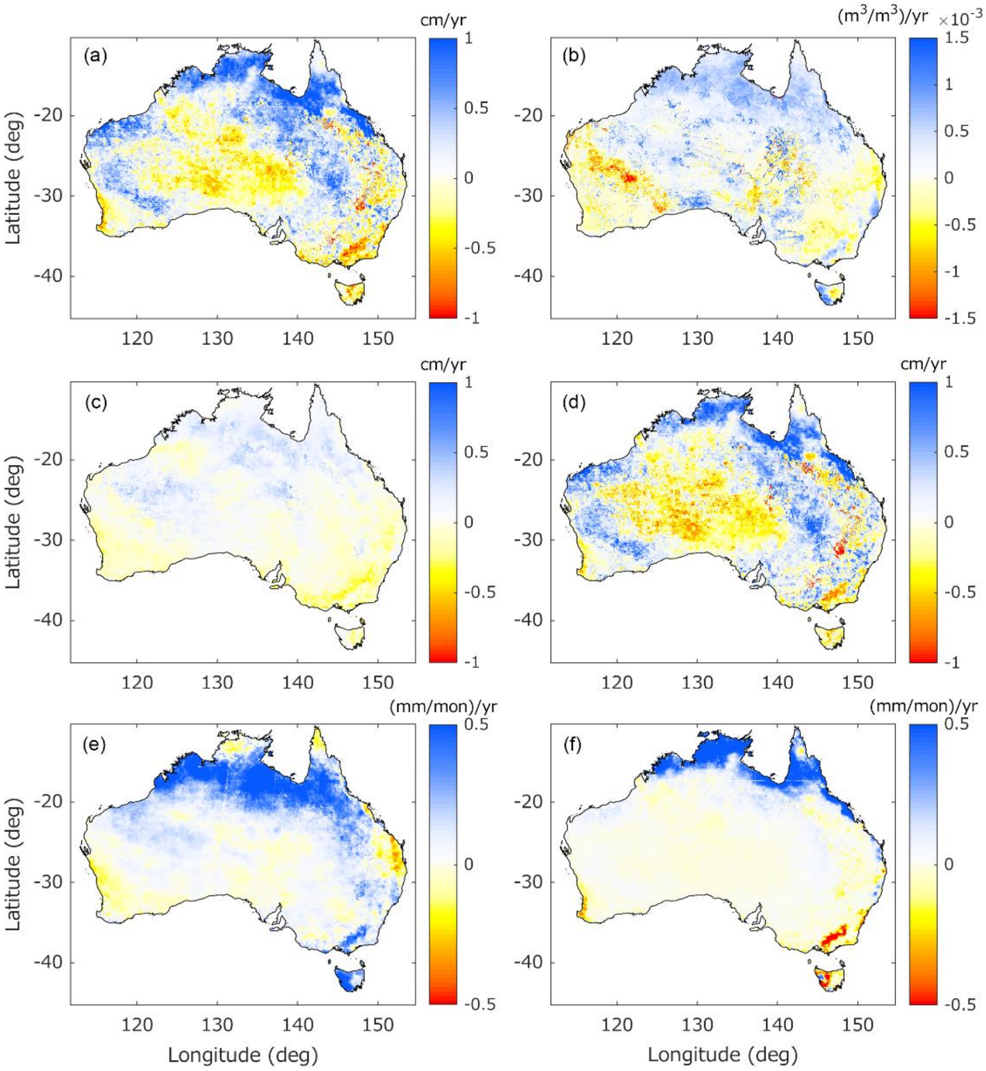
The correction can be found in **lines 452 – 454**.

C9: 780: Should it be Fig. 14?

R9: We again greatly thank Reviewer #1 for the correction. The typo is corrected to Fig. 14 (a-c).

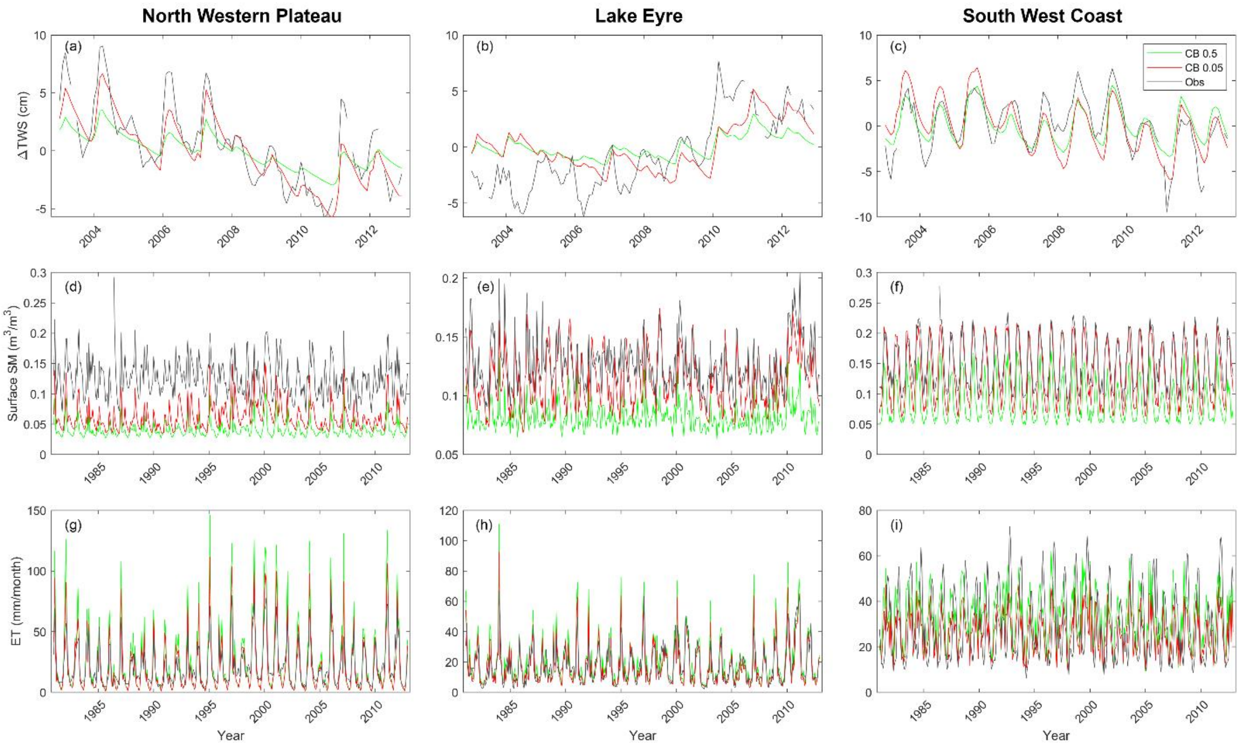
C10: 805: How about use the same color bar for Fig. 7c with Fig. 7a and d, so that the figure will show more details about the relationship and the comparison.

R10: We thank Reviewer #1 for the suggestion. The same color scale is used for variables with the same unit in the revised paper.



C11: 822: What is the unit of x axis for the delta TWS (the first row)? Please add the unit for the figures if necessary.

R11: The unit of the x-axis is year. For clarity, we change our x-axis to a full year. The label of the x-axis is also given.



C12: 823 - 824: Why not conduct the analysis over basins not for the whole Australia? Since you do the analysis for the whole area in the previous analysis.

R12: We believe Reviewer #1 means "why conduct the analysis over basins not for the whole Australia". Please note that the analysis over the entire Australia is already shown in Fig. 10 (as Reviewer #1 mentioned). We analyze the basin-average time series here since the characteristics of the estimate variables are not uniform across Australia. Displaying the average time series of all Australia leads to the same conclusion but omits small temporal details that can only be seen at the basin level.

We finally would like to thank Reviewer #1 again for taking the valuable time to review our manuscript. We hope that our responses clarify Reviewer #1 concerns. All suggestions will be implemented in our revised manuscript.

Response to Reviewer #2

The authors presented a well-written paper about a high-quality study in which they increased CABLE LSM resolution and assimilated the GRACE data into the model to improve its accuracy. Overall, I am satisfied with the quality of the paper. Some minor revisions can help to increase the quality of the paper. Below you can find my suggestions in this regard.

We greatly appreciate Reviewer #2's positive feedbacks and constructive comments. We provided our responses (R, in blue) to each of the reviewer's comments (C, in black) below:

C1: 1) Line 45 and Line 115 LSM model: You used CABLE LSM in your study. What is the advantage of CABLE over other LSMs? Apart from being frequently used in Australia (your study area), what are other reasons/motivations for choosing CABLE as the main LSM in your study?

R1: We thank Reviewer #2 for addressing our unclear explanation. Please note that CABLE is a global model and has been used to simulate global storage and fluxes (e.g., Decker et al., 2015; Haverd et al., 2018). This study uses Australia as a case study due to the availability of in situ data for validation. Despite the small community size, CABLE has been updated regularly by the CABLE community to catch up with the state-of-the-art development (e.g., Decker et al., 2015; Heverd et al., 2018). However, the current 0.5°×0.5° model spatial resolution limits its application to TWS studies in large river basins, and the attempt to improve CABLE spatial details has not been considered thus far. This is in contrast to other models, in which a high-resolution version has already been developed. Our development of CABLE 0.05° aims to narrow this gap.

To clarity this, we rewrite our opening statement (in the introduction section) as follows (see **lines 75 – 83**):

The Community Atmosphere Biosphere Land Exchange (CABLE; Kowalczyk et al., 2006) is an open-source global LSM developed and updated by the community. CABLE is a core LSM of the Australian Community Climate and Earth System Simulator (ACCESS, Bi et al., 2013; Kowalczyk et al., 2013) that can be used to simulate water storage and fluxes globally. The model has been regularly updated to incorporate the state-of-the-art model physics (e.g., Decker et al. 2015; Ukkola et al., 2016; Heverd et al. 2018). Despite its success, CABLE's spatial scale is currently limited to 0.5° (~50 km) due to the 0.5° resolution of its parameters and forcing datasets. This contrasts with other global model developments, where high-resolution versions have already been developed (e.g., van Dijk et al., 2013; Sutanudjaja et al., 2018). CABLE and its inputs must be reconfigured to increase the spatial detail of TWS estimates for smaller-scale studies (e.g., 0.01° – 0.05°). Our effort to increase the regional or local study's spatial resolution should narrow this development gap and has not previously been implemented.

C2: 2) Line 141 and section 3.2.1: Resampling coarse data to high resolution, bears extra uncertainty. How do you deal with this additional uncertainty?

R2: The effect of up/downscaling is also included in our DA process. In our perturbation process, when the data are resampled, their errors are also adjusted based on an error propagation approach. The relationship between coarse and fine-scale error can be expressed as:

$$\sigma_c = \frac{1}{M}\sum_{h=1}^{M}\sum_{l=1}^{M}\sqrt{\sigma_{f\ hl}^{2}\exp\left(\frac{-\phi_{hl}^{2}}{2\phi_{0}^{2}}\right)}, \tag{1}$$

where σ_c and σ_f represents a coarse and fine-scale error, (h, l) is the index of a grid cell, M is the number of fine-scale grid cells used in resampling, ϕ is a spherical distance between grid cells, and ϕ_0 is the considered correlation length (e.g., a coarse-scale's grid size).

We understand that this error size might not perfectly represent the truth (which is unknown), but it represents a more realistic error that changes with the increased/decreased spatial resolution. For clarity, we will add the above explanation to Sect. 3.1.

In Sect. 3.2.1, the resample is applied only to overlay the model grid cells with satellite products before comparison. We did not perform an error analysis here. Please note that remote sensing products may also contain bias and do not necessarily represent the truth. The inter-comparison is not used as validation. It is only used to assess the consistency between two independent estimates. The statement can be found in **lines 376 – 379**:

… the remote sensing products may contain biases (caused by, e.g., background model, processing algorithm) and do not necessarily represent the truth. (Ground truth validation is performed in Sect. 4.3.) The inter-comparison performed in this section is only to assess the consistency between two independent estimates: model and satellite…

C3: 3) Section 3.1: Why did you choose 3D Ensemble Kalman Smoother in your study and did not choose other DA/smoother methods? Please provide few lines about the benefits and potential limitations of this DA method. Also, for completeness, please discuss briefly why a Smoother can be a better choice than a Filter in your study?

R3: The 3D EnKS is chosen for two reasons. One, it accounts for spatial correlations in model errors and observation errors. The latter are highly correlated at neighboring 0.5° × 0.5° or 0.05° × 0.05° grid cells. Two, EnKS does not require interpolation of the observations (as in Ensemble Kalman Filter (EnKF); Tangdamrongsub et al., 2015) and mitigates the spurious jump in water storage estimates caused by applying the updates at the end of the month only. The additional computational cost is small: handling large covariance matrices and running the model twice for each month. For clarity, the above explanation is added to **lines 209 – 215** of the revised paper.

C4: 4) Section 4.2: Please provide your reasoning/hypothesis on why GRACE DA impacts deep water storage more than other components of the utilized LSM. You stated that similar finding was reported in other studies. Do you see a similarity between those LSMs and CABLE that resulted in a similar impact of GRACE DA? Is this a location-specific finding (Australia) or can it be generalized to other regions?

R4: We thank Reviewer #2 for the suggestion. The gravity satellite (like GRACE) is sensitive to the Earth's mass variation, which is more significant in the deeper layer, e.g., groundwater. This explains the GRACE effectiveness in capturing low-frequency signals (e.g., long-term trend of GWS) seen in most GRACE literature (and this study). For clarity, we include the above explanation in our revised paper, **lines 423 – 425**:

… It dominates the entire water column in several basins (e.g., Indian Ocean, Lake Eyre, North West Plateau, South West Plateau). This behavior reflects the nature of GRACE: groundwater provides a majority of the seasonal changes to terrestrial water mass.

The similarity seen in the previous studies is the impact of GRACE DA on GWS estimates (e.g., a significant change in GWS caused by GRACE DA). We find that our statement might be unclear. As such, we revised our sentence to (see **lines 425 – 426**):

GRACE DA has been shown to significantly affect GWS in previous studies, e.g., Girotto et al. (2016), Tangdamrongsub et al. (2018), and Li et al. (2019).

The impact of GRACE DA on GWS was seen globally, e.g., Rhine River basin (Tangdamrongsub et al., 2015), Continental United States (e.g., Kumar et al., 2016), China (Tangdamrongsub et al., 2017), Australia (Tian et al., 2017), global (Li et al., 2019). Recent work by Yin et al. (2020) assimilated GRACE data into CABLE in North China Plain and also reported GRACE DA's positive impact of GWS. These studies confirmed a clear benefit of GRACE DA despite the different models used, DA configurations, or locations. However, in our case, it is difficult to make a conclusive comment on GRACE DA behavior globally based only on the results of our early development. For clarity, we will add a remark regarding the simulation performance in other regions. The following sentences are added to the conclusion section, **lines 485 – 492**.

This means TWS estimates can be reproduced with more spatial detail by CABLE 0.05 at locations outside the area studied here since high resolution forcing data and model parameters are available globally (or near globally). However, the performance of such simulations might differ from this study due to the uncertainty in model parameters and forcing data that vary with geolocations (e.g., Herold et al., 2017; Tifafi et al., 2018). This remark also applies to the performance of GRACE DA. Although the improvement of assimilating GRACE into CABLE is also seen in other regions, e.g., North-East China (Yin et al., 2020), it is still difficult to quantify the benefit of GRACE DA over global river basins based on these early developments of CABLE/GRACE DA. Validation is highly encouraged to ascertain the accuracy of TWS estimates when performing the simulation in other regions.

Additional reference:

Yin, W., Han, S.-C., Zheng, W., Yeo, I.-Y., Hu, L., Tangdamrongsub, N. and Ghobadi-Far, K.: Improved water storage estimates within the North China Plain by assimilating GRACE data into the CABLE model, J. Hydrol., 590, 125348, https://doi.org/10.1016/j.jhydrol.2020.125348, 2020.

C5: Earlier on line 78, you state " GRACE DA has shown positive impacts on most TWS components, including groundwater (e.g., Girotto et al., 2017; Nie et al., 2019), soil moisture (Jung et al., 2019), and snow (Kumar et al., 2016)." How do you reconcile this to what you found in your study.

R5: We thank Reviewer #2 for the suggestion. This discussion is added to our results section, **lines 454 – 455** and **lines 460 – 454**:

… Despite different study areas, LSMs, and validation data, our finding is in line with, e.g., Girotto et al. (2017) and Nie et al. (2019), who also found a significant impact of GRACE DA on GWS components.

… GRACE is sensitive to the low-frequency variation (originated from deeper stores) and cannot effectively capture SSM, which is dominated by a high-frequency signal (e.g., precipitation). As a result, GRACE DA is found to have a minor (or negative) impact on the top soil component in most GRACE DA studies (e.g., Li et al., 2012; Tian et al., 2017; Tangdamrongsub et al., 2020). The small impact on SSM estimates also agrees with Jung et al. (2019), who observed GRACE DA's small (or negative) impact over dry regions in West Africa.

C6: 5) Section Conclusion: What is the role of uncertainty of the CABLE model inputs on the DA results? In another word, if you used other public resources as the CABLE inputs, would you get different results out of DA-based models? You mentioned 250-m resolution SoilGrids data for future use. Based on my personal experience, I found SoilGrids data not very accurate in many locations. Why do you think by using this data you can improve your model? I would suggest that you discuss about it in the paper.

R6: We thank Reviewer #2 for the comment. The role of uncertainty in model inputs will be discussed (please see below). We also thank Reviewer #2 for addressing the soil map accuracy. We found that our wording might be unclear. We present the improved spatial details (not accuracy) in this context. However, we agree with Reviewer #2 that the impact of forcing data and parameter accuracy on model performance needs to be discussed. As such, we rewrite our conclusion as follows, see **lines 483 – 497**:

The enhanced CABLE model resolution developed in this study relies on improved parameter and forcing data. The land surface physics remains unchanged. The workflow can be adopted for other CABLE repositories or different LSM with only slight modifications, e.g., number of soil or vegetation types. This means TWS estimates can be reproduced with more spatial detail by CABLE 0.05° at locations outside the area studied here, since high resolution forcing data and model parameters are available globally (or near globally). However, the performance of such simulations might differ from this study due to the uncertainty in model parameters and forcing data that vary with geolocations (e.g., Herold et al., 2017; Tifafi et al., 2018). This remark also applies to the performance of GRACE DA. Although the improvement of assimilating GRACE into CABLE is also seen in other regions, e.g., North-East China (Yin et al., 2020), it is still difficult to quantify the benefit of GRACE DA over global river basins based on these early developments of CABLE/GRACE DA. Validation is highly encouraged to ascertain the accuracy of TWS estimates when performing the simulation in other regions.

Our development is only demonstrated between 1981 – 2012 due to the availability of the Princeton forcing data. Future development can consider extending the temporal record or further increasing the spatial resolution of TWS estimates. The timespan extension is feasible using reanalysis forcing data from the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2; Gelaro et al., 2017). Despite a slightly coarser spatial resolution than the Princeton data, MERRA2 datasets would allow TWS simulations to be extended to the near present.

Note that we suggested a possible improvement of spatial resolution (not accuracy) using higher spatial-resolution data in our submitted manuscript. However, our suggestion might be too optimistic because sub-kilometer global forcing data needed for model simulation are not currently available. As such, we remove the statement regarding a sub-kilometer resolution to avoid confusion.

Additional references:

Herold, N., Behrangi, A. and Alexander, L. V.: Large uncertainties in observed daily precipitation extremes over land, J. Geophys. Res.: Atmospheres, 122(2), 668–681, https://doi.org/10.1002/2016JD025842, 2017.

Tifafi, M., Guenet, B. and Hatté, C.: Large Differences in Global and Regional Total Soil Carbon Stock Estimates Based on SoilGrids, HWSD, and NCSCD: Intercomparison and Evaluation Based on Field Data From USA, England, Wales, and France, Global Biogeochem. Cycles, 32(1), 42–56, https://doi.org/10.1002/2017GB005678, 2018.

C7: 6) Code availability: It would be more useful to the readers if you could share your code for the DA framework.

R7: We thank Reviewer #2 for the suggestion and particular interest in the software. Software development is already on the list of our research plans. Despite our very limited resources, we are trying hard to make the software available as soon as we can.

C8: 7) Title of the paper: GRACE Data Assimilation implies a DA method that is called GRACE. To avoid confusion for readers who are not familiar with GRACE mission, I would suggest using “Assimilation of GRACE data” instead.

R8: We agree with Reviewer #2. We modify the title to “Development and evaluation of 0.05° terrestrial water storage estimates using CABLE land surface model and assimilation of GRACE data” as Reviewer #2 suggested.

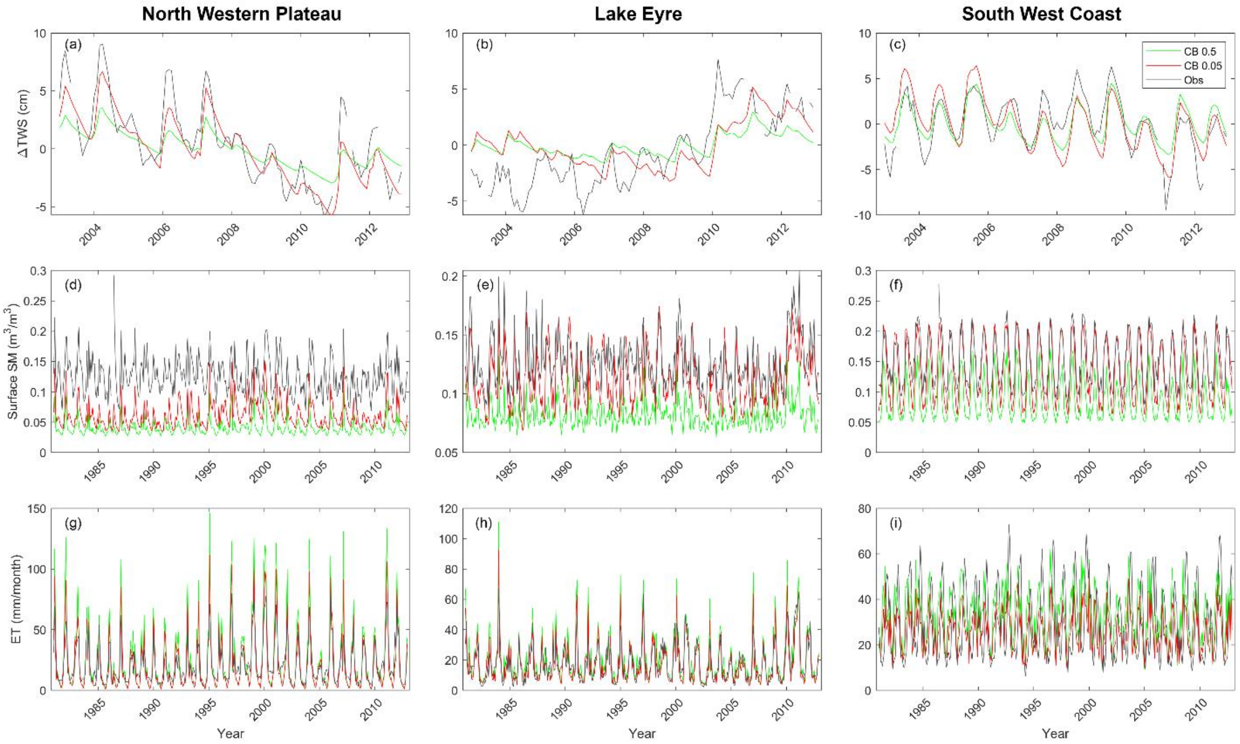
C9: 8) Table 2: grid size of Harmonized World Soil Data base is 30 arc-second ~ 0.0083 deg Please carefully check the rest of the data in this table.

R9: We greatly appreciate Reviewer #2 for pointing this out. This is a typo. We correct the error and carefully recheck the entire table. In our revised version, we express the grid size using native resolutions of the products for consistency with product descriptions.

	Products	*Grid size*	*Time interval*	*References*
Meteorological Forcing data	*Princeton forcing data version 2*	*0.5°*	*3 hours*	*Sheffield et al. (2006)*
Precipitation	*Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS)*	*0.05°*	*1 day*	*Funk et al. (2015)*
Soil type	*Harmonized World Soil Database version 1.2*	*30 arc-second*	*n/a*	*Nachtergaele et al. (2009)*
Vegetation type	*MODIS Land Cover Maps*	*500 m*	*n/a*	*Broxton et al. (2014)*
LAI	*Global Land Surface Satellite (GLASS)*	*0.05°*	*~8 days*	*Xiao et al. (2013)*
GRACE	*NASA GSFC Mascons*	*Irregular*	*~1 month*	*Luthcke et al. (2013)*

C10: 9) Figure 11 & 12: Please specify the label for x axes.

R10: We modify our x-axis to a full year. The label of the x-axis is also given. We show the modified version of Fig. 11 below, and the same correction is also applied to Fig. 12.



Finally, we would like to thank Reviewer #2 for taking the valuable time to review our manuscript. We hope that our responses clarify Reviewer #2 questions. All suggestions will be implemented in our revised manuscript.

Response to Reviewer #3

The manuscript presents a high resolution model (0.05°) for Terrestrial Water Storage (TWS). The model implemented (CABLE SubgridSoil GroundWater), that was previously used to estimated TWS at 0.5°, is 'upgraded' to 0.05° resolution and extended with GRACE satellite observations via Ensamble Kalman Smoother. The method is demonstrated on Australia, a complex case study, with different climatological regions. Processing of the 32-year time span on continental scale is an impressive test case.

We thank Reviewer #3 for the feedback and constructive comments. We provided our responses (R, in blue) to each of the reviewer's comments (C, in black) below:

C1: The introduction highlights the importance of high resolution TWS modeling for risk management (l. 31), however, the temporal range of 1981-2012 does not reflect this application. The conclusion states that this would have been possible with different data (l. 461), at even higher resolution (l. 463). Why wasn't this done?

R1: We agree with Reviewer #3 that our simulation period demonstrated here might not be realistic for the (real-time) risk management application. However, please note here that our statement in this context describes the importance of improved spatial resolution, see **lines 32 – 33**:

At a regional or local scale, the spatial resolution of the TWS estimate is vital, as most applications (e.g., risk management for drought or flood) require accurate information at the county or sub-county level (Quiring, 2009).

Our simulation was based on public global-datasets that are only available from 1980 to 2012. As our conclusion suggests, the new datasets can be used if they are available. Unfortunately, at the time we conducted our analysis, there were no comprehensive global forcing datasets with higher spatial resolution than 0.5°/0.05° AND time span longer than the 1981-2012 period. In conclusion, we suggest the possibility of using a slightly coarser spatial resolution forcing data (MERRA2) to extend the dataset to the near present. To clarify Reviewer #3 concerns, we include the above explanation in our conclusion, **lines 493 – 497**:

Our development is only demonstrated between 1981 – 2012 due to the availability of the Princeton forcing data. Future development can consider extending the temporal record of TWS estimates. The timespan extension is feasible using reanalysis forcing data from the Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2; Gelaro et al., 2017). Despite a slightly coarser spatial resolution than the Princeton data MERRA2 datasets would allow TWS simulations to be extended to the near present.

In the submitted manuscript, we also suggest a possible improvement of spatial resolution using higher spatial resolution data. However, we find that our suggestion might be too optimistic because sub-kilometer global forcing data needed for model simulation are not currently available. As such, we remove the statement regarding a sub-kilometer resolution to avoid confusion.

C2: The open character of the model (l. 81) is an important characteristic that deserves more attention, due to its high potential. (N.B. other reviewers signaled that the public code may not include the GRACE data assimilation, this is unclear to me from the text.)

R2: We thank Reviewer #3 for the comment. Please note that our statement here only describes the public datasets (e.g., forcing data, parameters, remote sensing data), not codes, see **lines 84 – 85**:

Our approach utilizes only publicly available global datasets, so resulting TWS estimates can be reproduced over any target region…

The GRACE DA approach developed in our study is thoroughly described in Sect. 3.1, and it can be simply reproduced using any computer language. However, we understand the need for GRACE DA software. Software development is already on the list of our research plans. Despite our very limited resources, we are trying hard to make the software available as soon as we can.

C3: l. 56-64 list various models of comparable spatial resolution.

R3: We thank Reviewer #3 for pointing this out. We reported the native unit of models/products and did not convert them to avoid rounding errors. To clarify Reviewer #3 concern, we report the native resolution of the models/products and include km or degree unit (or both) when they are available.

C4: l. 76 "GRACE DA has shown positive impacts […]", combined with l. 90 "[t]he objectives of this paper are […] 2) to assess the GRACE DA impact on […] CABLE […]". What does CABLE provide that previous studies did not?

R4: We thank Reviewer #3 for the suggestion. For clarity, we revise our introduction as follows, see **lines 90 – 97**:

Recent studies have shown success in assimilating GRACE data into a coarse-scale CABLE version to improve TWS and groundwater storage (GWS) estimates in the Goulburn River catchment and in the North China Plain (Tangdamrongsub et al., 2020; Yin et al., 2020). In this study, GRACE observations (Luthcke et al., 2013) are also assimilated into CABLE 0.05° (and CABLE 0.5°) to improve the accuracy of TWS components between 2003 and 2012. Assimilating the coarse GRACE observations into a much higher-resolution model is performed using the 3-dimension ensemble Kalman smoother (EnKS 3D; Tangdamrongsub et al., 2017). This approach will reveal whether assimilating GRACE data can benefit a newly developed fine-scale CABLE configuration. Our study will perform a thorough investigation on this issue to address GRACE DA's benefit on CABLE 0.05°.

C5: The GRACE data set spans only a small part of the 1981-2012 time span of the study. How is GRACE data integrated, outside the periods of data assimilation?

R5: We thank Reviewer #3 for addressing this question. GRACE data is only assimilated between 2003 and 2012 as clearly explained in **lines 404 – 406**:

GRACE observations are assimilated into the CABLE 0.5° and CABLE 0.05° models (called GRACE DA 0.5° and GRACE DA 0.05°, respectively) between January 2003 and December 2012 (due to the availability of meteorological forcing and GRACE data).

GRACE DA is not performed when data are not available, e.g., prior to 2003. The DA evaluation is only performed in 2003-2012 period. To clarify this further, we also include the GRACE assimilation period in Sect. 2.3, **lines 173 – 174**:

In this study, GRACE data are assimilated into CABLE between January 2003 and December 2012 (due to the availability GRACE data).

In addition, our data processing diagram (Fig. 3) also clearly explains that GRACE data are assimilated only when they are available. Please note that the flowchart is already modified based on the Reviewer #3 suggestion (please see R8).

C6: I would suggest to report both input and validation/evaluation data sets (§ 2.4) in similar fashion. For example, include the evaluation (satellite) data in a similar fashion in Table 2.

R6: We thank Reviewer #3 for the suggestion. The characteristics of evaluation (satellite) data are included in Table 2 as Reviewer #3 suggested:

	Products	*Grid size*	*Time interval*	*References*
Meteorological Forcing data	*Princeton forcing data version 2*	*0.5°*	*3 hours*	*Sheffield et al. (2006)*
Precipitation	*Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS)*	*0.05°*	*1 day*	*Funk et al. (2015)*
Soil type	*Harmonized World Soil Database version 1.2*	*30 arc-second*	*n/a*	*Nachtergaele et al. (2009)*
Vegetation type	*MODIS Land Cover Maps*	*500 m*	*n/a*	*Broxton et al. (2014)*
LAI	*Global Land Surface Satellite (GLASS)*	*0.05°*	*~8 days*	*Xiao et al. (2013)*
TWS	*GRACE NASA GSFC Mascons*	*Irregular*	*~1 month*	*Luthcke et al. (2013)*
Soil moisture	*European Space Agency - Climate Change Initiative program (ESA-CCI)*	*0.25°*	*1 day*	*Dorigo et al. (2017)*
Evapotranspiration	*Global Land Evaporation Amsterdam Model (GLEAM)*	*0.25°*	*1 day*	*Martens et al. (2017)*

C7: Upsampling of precipitation (l. 144) may not reflect natural precipitation patterns. Likewise, nearest neighbor interpolation (l. 142) may introduce strong gradients.

R7: We agree with Reviewer #3 that the resampled precipitation might not reflect its natural patterns. We are aware of this additional error and treated it statistically. The effect of up/downscaling is also included in our DA process. In our perturbation process, when the data are resampled, their errors are also adjusted based on an error propagation approach. The relationship between coarse and fine-scale error can be expressed as:

$$\sigma_c = \frac{1}{M}\sum_{h=1}^{M}\sum_{l=1}^{M}\sqrt{\sigma_{f\;\;hl}^{2}\exp\left(\frac{-\phi_{hl}^{2}}{2\phi_0^{2}}\right)}, \tag{1}$$

where σ_c and σ_f represents a coarse and fine-scale error, (h, l) is the index of a grid cell, M is the number of fine-scale grid cells used in resampling, ϕ is a spherical distance between grid cells, and ϕ_0 is the considered correlation length (e.g., a coarse-scale's grid size).

We understand that this error size might not perfectly represent the truth (which is unknown), but it represents a more realistic error that is changed with the increased/decreased spatial resolution. For clarity, we will add the above explanation to Sect. 3.1.

In addition, the impact of the resampled forcing data is also discussed in **lines 309 – 401**:

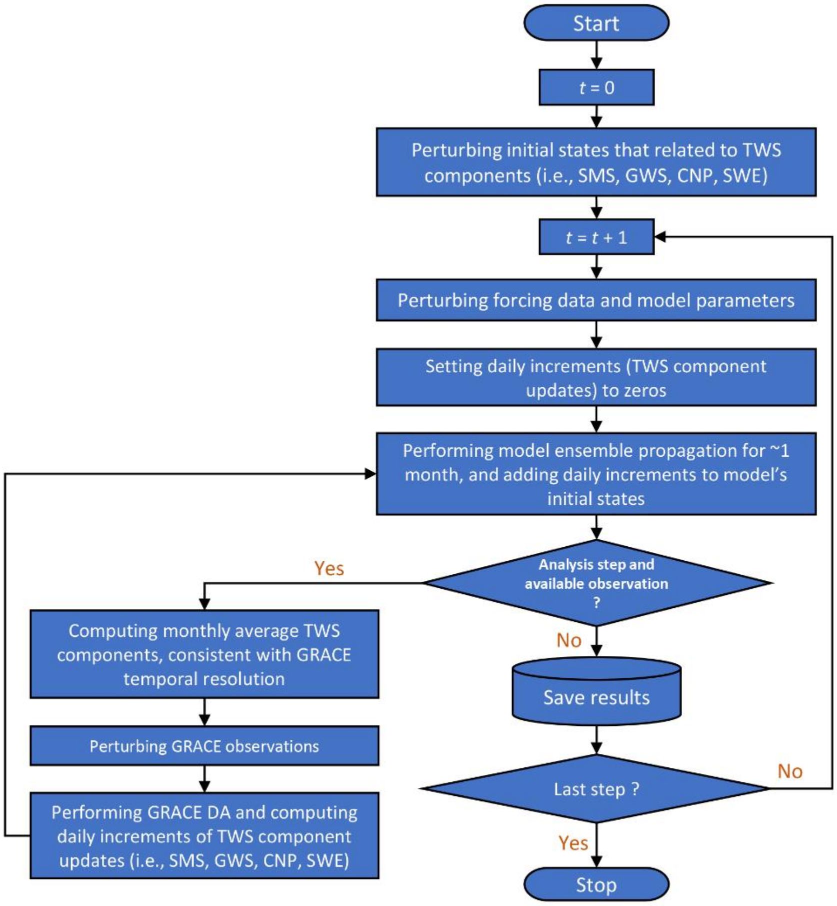
…The use of coarse resolution forcing data (e.g., precipitation) could also explain the small TWS amplitude observed in CABLE 0.5°. Coarse scale forcing data averages local precipitation signals over a larger area than does the finer resolution forcing data, resulting in a smaller amplitude.

C8: What is provided to CABLE, and what is optimized in the Ensemble Kalman Smoother. A 'data flowchart' would be a helpful extension to Figure 3.

R8: We thank Reviewer #3 for the suggestion. As described in **lines 231 – 232**, the updated state variables are six soil moisture layers, canopy storage, snow water equivalent, and groundwater storage:

The state vector consists of nine model states (n = 9): six soil moisture layers, canopy storage, snow water equivalent, and groundwater storage.

As Reviewer #3 suggested, we will update our processing diagram to include more processing details:



C9: l. 243 "[…], the daily increment (ΔAd) of the update is computed by dividing ΔA by the total number of days in that month." How does this influence high frequency signals (e.g. precipitation spikes)?

R9: Because GRACE by nature provides lower frequency observations, it is unlikely to capture signals at surface layers (e.g., top soil) that is governed by high-frequency signal (e.g., from precipitation). The applied increment in EnKS is to distribute the GRACE monthly update throughout the month. The limitation of GRACE DA on high-frequency signals is already discussed in the manuscript **lines 459 – 462**:

… The small impact is attributed to limited GRACE sensitivity. GRACE is sensitive to the low-frequency variation (originated from deeper stores) and cannot effectively capture SSM, which is dominated by a high-frequency signal (e.g., precipitation). As a result, GRACE DA is found to have a minor (or negative) impact on the top soil component …

C10: § 3.2.1 How is in-situ (point) data (§ 2.4.2) handled?

R10: The in situ data is not resampled. Our sensitivity analysis (not shown) reveals that performing the validation at in situ data location or at model grid cells leads to the same outcome. This is due to the fact that there are not many in situ data in the same model grid cell.

C11: § 3.2.3 What will this metric express?

R11: The regression is used to estimate the long-term trend and seasonal amplitude of TWS components. The results are discussed in Sect. 4.1.1, 4.1.2, and 4.2.

C12: Is there any indicator for the timing of the signal (e.g. systematically late detection)?

R12: The timing indicator is not used in our study. Our analysis is only based on correlation coefficient and RMS, consistent with all GRACE DA studies, e.g., Zaitchik et al. (2008), Tian et al. (2017), Kumar et al. (2016), Girotto et al. (2016).

C13: § 3.2.4 The term ‘spatial resolution’ is very confusing in the context of the study, see also l. 289. Also l. 450 “the 0.05° model also improves the spatial resolution by a factor of two to three over the 0.5° version,” is counter-intuitive.

R13: We thank Reviewer #3 for raising this concern. Please note that a spatial resolution is defined as a minimum distance at which two signals of equal magnitude can be separated. As it is well understood, the spatial resolution is not necessarily equal to the model grid size. In other words, the 0.05° estimated variable may not have 0.05° resolution. We understand that the terms spatial resolution and model grid size can be confusing. As such, we carefully explain how the spatial resolution can be determined in Sect. 3.2.4 and provide its schematic in Fig. 4. We also clearly explain the difference between spatial resolution and grid size in **lines 309 – 310**:

It is noteworthy that the 0.5° or 0.05° represents the CABLE grid size, which may differ from the spatial resolution. The term “spatial resolution” used in this paper refers to the determined resolution computed from Sect. 3.2.4.

For clarity, we add the definition of spatial resolution in the revised version, see **line 293**: “*Spatial resolution is defined as the minimum distance at which two signals of equal magnitude can be separated*”.

C14: There are no comments on the computational resources required by the model. Given the open character of the model, some hints would be welcome.

R14: The details on model source code, installation, and computational resource requirement can be found in the CABLE webpage (internet links are provided in Data availability section). CABLE is developed using Fortran and can be executed in a Unix environment. The input/output file format follows

NetCDF Climate and Forecast (CF) convention. We run the model in high-performance computing (HPC) environment. For clarity, we include the details of CABLE software in our revised version, **lines 124 – 125**:

CABLE is developed using Fortran and can be executed in a Unix environment. The input/output file format follows NetCDF Climate and Forecast (CF) convention…

C15: More overarching conclusions/summaries, for each comparison, would make the section more readable. What is the 'take home message' from each comparison.

R15: We thank Reviewer #3 for the comment. Please note that each analysis contains a summary sentence. For clarity, we will add an additional summary statement to the last sentence of our analysis (if it is not already there). Please also note the summary from each comparison can also be found in the conclusion section (in chronological order), **lines 473 – 480**:

This study enhances the spatial resolution and timespan (> 30 years) of regional TWS estimates using the CABLE LSM, high-resolution land cover maps and forcing data, and GRACE DA application. By improving the model parameter and forcing data (without GRACE DA), the developed CABLE 0.05° model shows clear improvements in the accuracy of water balance component estimates (e.g., soil moisture, groundwater, evapotranspiration) compared with in situ and independent satellite data. The 0.05° model also improves the spatial resolution by a factor of two to three over the 0.5° version. The extended timespan provides insightful information for long-term assessment of regional water resources and climate variability. The enhanced model parameterization is found to play a significant role in the improved TWS estimates. Incorporating GRACE DA into the model leads to further improvements of TWS component estimates. The positive impact of GRACE DA is found in the deep storage component (e.g., GWS), while the impact on the surface components and flux estimates (i.e., SSM and ET) is trivial. Of the four case studies investigated here, the most accurate simulation uses CABLE 0.05° with GRACE DA…

C16: Various spatial units are mixed together, making difficult to compare between models and sources. The resolution of this model is reported in degrees, while the resolution of relevant models are mentioned in kilometers (l. 56-64). A single unit would be best, or report both.

R16: We understand that the mixed unit in the same paper might be confusing. Our paper reports the native unit of models/products and did not convert them to avoid rounding errors. To clarify Reviewer #3 concern, we report the native resolution of the models/products and include km or degree unit (or both) when they are available.

C17: Ambiguous terminology, e.g.:

R17: Changes will be made. Please see below:

meaningful, l. 8, l. 33;

We will change meaningful resolution to high resolution.

sufficiently, l. 43, l. 52;

sufficiently will be removed from the sentence.

can, l. 44 (instead, provide a figure such as the longest available time series).

We thank Reviewer #3 for the suggestion. We will include two good references (e.g., Flechtner et al., 2014; Karthikeyan et al., 2017) where readers can obtain periods of available satellite observations.

Flechtner, F., Sneeuw, N. and Schuh, W.-D., Eds.: Observation of the System Earth from Space - CHAMP, GRACE, GOCE and future missions: GEOTECHNOLOGIEN Science Report No. 20, Springer-Verlag, Berlin Heidelberg., 2014.

Karthikeyan, L., Pan, M., Wanders, N., Kumar, D. N. and Wood, E. F.: Four decades of microwave satellite soil moisture observations: Part 2. Product validation and inter-satellite comparisons, Advances in Water Resources, 109, 236–252, https://doi.org/10.1016/j.advwatres.2017.09.010, 2017.

l. 30, "[…] spatial resolutions are relatively coarse […] models that primarily focus on global or continental scale";

The statement will be modified to (see **lines 30 – 31**):

.. spatial resolutions are coarse due to the limitation of sensors and models that focus on …

l. 190 "[…] distributed unevenly across the continent."

The statement will be modified to (see **line 198**):

… are distributed across the continent...

C18: Inline hyperlinks, except for the sources (l. 467 onward), make the text difficult to read, l. 60, l. 62, l. 114, l. 156, l. 161, l. 189.

R18: We thank Reviewer #3 for the comment. We will move the inline hyperlinks to Data availability sections.

C19: Table 1, l. 110, appears to be misplaced, l. 69. The reference in l. 106 could point back.

R19: We thank Reviewer #3 for the comment. Table 1 is placed after a paragraph when it is first mentioned, which is after Sect. 2.2 (at line 110).

C20: l. 130, "still"?

R20: "still" will be removed from the sentence. The sentence now reads (see **lines 138 – 139**):

The soil map is also derived from the Harmonized World Soil Database but at 0.05° grid spacing…

C21: l. 132, Figure 1 does not show the high resolution of the data. Maybe include a detail?

R21: We thank Reviewer #3 for the comment. Please note that only rederived parameters are shown in Fig. 1. These parameters are used in the CABLE 0.05° configuration. This is stated in **lines 140 – 141**:

All rederived parameters are shown in Fig. 1.

C22:l. 149, 320 years → 32 years

R22: Please note that the model is spinning up for 320 years to obtain the initial states. This is stated in **lines 158 – 159**:

initial states are obtained using a 320-year spinup, i.e., performing ten repeated runs between 1981 and 2012.

C23: l. 185, "Murray-Darling Basin", see Figure 1.

R23: We thank Reviewer #3 for the suggestion. The statement will be modified to (see **line 194**):

…Murray-Darling Basin (see Fig. 1).

C24: l. 250, § 2.3 should be § 2.4

R24: We thank Reviewer #3 for the correction. Change will be made.