Author's response

GMD manuscript: Coupling a large-scale hydrological model (CWatM) with a high-resolution groundwater flow model to assess the impact of irrigation at regional scale. *Guillaumot et al.*

According to editor comment, we modified the title :

Coupling a large-scale hydrological model (CWatM v1.1) with a high-resolution groundwater flow model (MODFLOW 6) to assess the impact of irrigation at regional scale

First reviewer

Thank you for these comments. We answered to each comment below in italic and labelled AR (Authors Response). Corrections to the manuscript are written in blue.

Specific comments

1) The landuse category "groundwater-supported grasslands" is problematic. It is defined through model runs and a sort of arbitrary threshold (4 out of 12 months), i.e, is not a property of "nature". The classification is obtained through initial model runs and then kept fixed – however, with optimal parameters, some of the cells could change their classification (the threshold crossed). Did you consider this feedback loop?

AR : Thanks for this relevant comment, we are convinced that it will improve the manuscript. Indeed, the area of the landuse category "groundwater-supported grasslands" need to be defined within each mesh of CWatM. The spatio-temporal extent of groundwater capillary rise (and the resulting contribution of groundwater to evapotranspiration) is an uncertain component of hydrology and there is no observed datasets providing "groundwater-supported lands" maps. Only hydrogeological models can provide an estimate of this process occurring where groundwater levels are shallow. The calibration has been done without including this specific land cover to avoid what the reviewer mentioned in the comment. Thus, we assume that adding this land cover does not impact the calibration (only locally). This landuse category was added to study more accurately the impact of irrigation. To clarify the manuscript, we move the description of the "groundwater-supported grasslands" (section 2.2.2) inside section 5 "Experiments to infer the impact of irrigation", now from line 370 to 372, and completed by:

"Therefore, choosing a threshold of "4 months out of 12" is a compromise allowing to focus on areas significantly supported by groundwater. This illustrates the importance of simulating well water table depths."

The threshold (4 out of 12 months) is arbitrary but a good compromise, so that for some months and some meshes, this land cover fraction is too small, but in this case the excess groundwater support is diluted within other land covers, while this land cover is too large for some months so that groundwater capillary rise is diluted.

2) The description of the management implementation into the model is rather vague. Most prominently, (l. 218) "outfitted with daily-reservoir-specific operations" and then "are set to satisfy some agricultural demands" (l. 222) are so unclear that no other modeler could reproduce your results. You should provide more details or a different way of presenting the inclusion of irrigation and pumping.

AR : Thank you for this remark, we bring several improvments to the model regarding water management: the implementation of groundwater pumping (mentionned previously) and canals. Water allocation was not problematic as we simply prioritized non-irrigation requests. Water sources was more challenging, as explained below. We refer to Smilovic et al. (2022) for further details. We modified the paragraph line 200-207 in order to clarify:
"Note that here, we benefit from a new CWatM development, including surface water and agricultural management (discussed further by Smilovic et al., 2022). To appreciate the impacts of surface water management on groundwater, more than 40 existing reservoirs are considered in the Bhima basin. They are simulated with daily reservoir-specific operations and connected to pre-defined specific spatial distribution areas (command areas provided by local experts) through canal networks. Therefore, reservoirs in the model distribute water based on daily command area demand and according to daily maximums with preference given to non-irrigation requests (domestic, industrial, and livestock). A fraction of this water leaks through the canal network. Further, rivers and lakes are set to satisfy up to 20% of cell-specific agricultural demands depending on availability. Water demand is supplemented with groundwater when surface water volume does not coincide with the need."

3) The parameter optimization (to obtain values for aquifer thickness, permeability and porosity) is not described to any detail. It was done as an inverse-modelling approach "calibrated manually by comparing simulated and observed water table…" (l. 348). How many boreholes did you use for calibration, and how many were then used for validation? How certain are you that the simple values obtained for Theta and T are the optimal ones (in the sense of minimum C_mean? Or nRMSE? Or both?) In any case, the values obtained are model-internal, needed to obtain sufficient model performance, not properties of the basins; one aspect of this mismatch is that all three are local properties, spatially varying, whereas your model approach is ignoring this heterogeneity. It is therefore also not justified to claim that "the GLHYMPS database overestimates these values" – no, a valid statement would be that your optimal parameters obtained in the CWatm-MODFLOW approach are smaller than those from the database.
AR : We completed several parts to clarify the parameter optimization following these comments. The number of boreholes is given just above in section 4.1 « Available observed data ». In the new version, we also mention it in section 4.2 « Comparison between observed and simulated water table » dealing with parameter optimization (line 310-311):
"As described above, the calibration relies on 62 and 351 boreholes for the Seewinkel and Bhima, respectively."
Aquifer thickness is not optimized and was determined in consultation with groundwater experts. In Bhima, aquifer thickness was determined to be 50 m to reflect the depth of functional storage, determined in consultation with local experts, and not included in the calibration. Hydraulic conductivity and specific yield variables were included in the calibration with multiple discharge stations, and ranges for calibration were derived from Surinaidu et al. (2013). The top 20% from the discharge calibration were further analyses for water table fluctuation and depths. This is now included line 335-341:
"For Bhima, hydraulic conductivity and specific yield variables were included first in a calibration (Fortin et al., 2012) using five daily discharge stations (2000-2009), and ranges for calibration were derived from Surinaidu et al. (2013). The top 20% from the discharge calibration were further analyzed for water table fluctuation and depths. Groundwater parameters hydraulic conductivity and porosity are further calibrated by comparing simulated and observed water table recorded monthly in boreholes from 1983 to 2016 and twice a year from 1997 to 2009, respectively, for the Seewinkel and Bhima. Standard CWatM parameters were used in Seewinkel. Therefore, only groundwater parameters are calibrated in this study."
For given pumping and recharge timeseries provided by CWatM, transmissivity (hydraulic conductivity times saturated thickness) first constrains Cmean. Then, once transmissivity is defined, porosity can be constrained by the nRMSE criterion focusing on water table fluctuations. Thus, we further calibrated hydraulic conductivity and porosity by trial-errors. Our approach consists in

simulating hydrologic systems at large scale with a physically-based representation of groundwater. In this context, we tried to find equivalent homogeneous properties to model the system based on water table observations, our results suggest that we can ignore heterogeneity at the aquifer scale (as highlighted by Houben et al., 2022) even if it would lead to inaccurate results locally such as under/over-estimated water table depth locally. Moreover, obtained parameters are compared to sparse data in section 6.1. This point appears in Discussion (section 7.2, line 529-532):
"We infer that, by calibrating water table fluctuations separately, we simulated a good combination between groundwater recharge, pumping, and aquifer lateral flow (driven by hydraulic diffusivity). Therefore, water table time fluctuations contain important information (Houben et al., 2022), even when reproducing the absolute water table depth remains challenging, particularly in Bhima, as also noted by Vergnes et al. (2020)."
We agree in part with the comment regarding the comparison with the GLHYMPS database. This is a reference for hydraulic conductivity and porosity at global scale. However, the database itself contains a lot of uncertainty as it extends to the globe a correlation between hydraulic conductivity and geology based on existing calibrated hydrogeological models in several regions. The aim of section 6.1 was to compare our parameters with reference local data and also to compare them with large-scale datasets usually used. We corrected the section 6.1 (line 375-383):
"Optimal aquifer permeability and porosity (Table 1) are compared to different datasets as described below. The permeability and porosity in Seewinkel are 5×10^{-5} m/s and 0.07, respectively. These values are respectively 3.10^{-4} m/s and 0.19 in the global GLHYMPS database (Gleeson et al., 2014; Huscroft et al., 2018). The imposed aquifer thickness (20 m) agrees with the global depth-to-bedrock map (Shangguan et al., 2016), where the average value for the study area is around 22 m.
In Bhima, the permeability and porosity are 1.2×10^{-5} m/s and 0.018, respectively. These values are respectively 3.16×10^{-6} m/s and 0.09 in the global GLHYMPS database. The global depth-to-bedrock values consider thinner aquifers in this region (from zero to a few meters) compared to the imposed aquifer thickness (50 m). Note that for similar permeability (1×10^{-5} m/s) and thickness (50 m), Surinaidu et al. (2013) calibrated a porosity of 0.01–0.03 in a regional model. This suggests that permeability is underestimated and porosity overestimated in GLHYMPS over this basin."

4) The quality assessment of the model is positively biased due to subjective choices of the modelers: first, boreholes at the edges are discarded since lateral flow is not properly implemented in the model; then, incomplete records (less than 50% values available) are discarded, without obvious reason; then, and worst, 5% "bad" boreholes (with particular poor model performance) are also excluded. The remaining time series have a higher performance by construction, but exactly how much better? Also, it is wrong that the locations where the model is performing particular poor are "impacted by specific local conditions", you can't know that since there are no independent validation data? Please elaborate on this; the reported C_mean, nRMSE and KGE values are necessarily biased.
AR : Quality assessment is very slightly biased as we keep 76 and 94 % of monitoring boreholes. The aim of this step is to infer equivalent hydraulic conductivity and porosity from sparse water table observations so that the overall behavior of the aquifer is well reproduced. Thus, quality assessment requires to pre-process observed data to exclude the less representative boreholes (or boreholes where the model will never reproduce water table anyway). Water table time fluctuations can not be calibrated if we do not have enough time data, several boreholes are excluded for this reason. If not some wells would be calibrated only during one or two seasons, or worst, only during humid or dry periods. We acknowledge that water table depths at the edges of the basin are surely less realistic because the 'no flow' boundary condition considers that the topographic limit of the basin is similar with the hydrogeological limit. This effect is important only for meshes at the edges of the basin and does not impact the simulation because the basin is huge. So, we think it makes sense to not consider these boreholes. Removing 5% "bad" boreholes is more critical but we can exclude manually, one by one, boreholes presenting bad quality data, and investigating why they have a different behavior or measurement errors. We added line 302-304:

"Indeed, because model assumptions such as homogeneous aquifer or pumping wells location, it can be inferred that some boreholes could not be well represented by the model due to the vicinity of pumping wells, rivers or local heterogeneities."

Technical corrections

l. 31: "subgrid resolution" is something else as "river incision", the parenthesis seems to indicate that the latter is an explanation of the former, which is not the case. Please correct.
AR : Yes indeed, thank you for this remark. In the new version, we modified the sentence by (l. 33-34, abstract) : "We found that grid resolution is the main factor that affects water table depth bias because it smooths river incision, while pumping affects time fluctuations."

l. 52: why should „unrealistic aquifer properties" always UNDERestimate water table depth? It could also be the other way round. Please correct or explain.
AR : Indeed, an explanation is missing and the sentence can be misunderstood, thanks. We modified by (line 53-55) : "In addition, coarse resolutions of groundwater representation tend to smooth hydraulic gradients, leading to unrealistic aquifer properties (Shrestha et al., 2018) and potentially to underestimate water table depth drawdowns due to groundwater pumping because withdrawals are applied to entire coarse grid cells instead to be applied to punctual boreholes."

l. 91: Please expand on "environmental flow limit", at least provide a reference here.
AR : A first reference is already mentioned just before (De Graaf et al., 2019). To clarify, we completed the new version by (line 93-95) : "They identified that rivers reach their environmental flow limit (meaning that groundwater baseflow supporting rivers fails below its 10th percentile as suggested by Gleeson and Richter, 2018) before substantial groundwater depletion occurs".

l. 303: no, this information is not contained in Table 1. There is also no Table 2, so what do you refer to here?
AR : Thanks, we modified the beginning of this paragraph to clarify (line 285-287) : "Table 1 provides mean simulated water withdrawals. Annual surface and groundwater withdrawals within the Bhima basin are estimated for different sectors for 2013 in the Upper Bhima Subbasin draft report. The model simulates these withdrawals closely (5% higher), with higher groundwater use and lower surface water use."

l. 306: out from the deficit of previous reports for the Bhima basin (e.g. canal leakage is not included), it can't be concluded that your model is appropriately representing the region's annual water use. The "therefore" is a logical fallacy.
AR : The model simulates too much groundwater withdrawals and not enough surface withdrawals compared to estimated data. Our point is that estimated groundwater pumping are underestimated because of under-reporting and because the reports underestimate groundwater availability as they did not include groundwater recharge due to canal leakage. Excluding canal leakage also contributes to overestimate surface water withdrawals. So, we can conclude that the model well represents annual surface and ground water use. We modified the last sentence to clarify (line 290-292) : "In agreement with the uncertainty of the reported values and following experts opinion mentioned above, this model appropriately represents the region's annual surface and ground water use."

l. 327: how can topography "extend over several orders of magnitude" (and of what variable? Slope?)?

AR : Our point is that large scale models usually compare absolute water table. Because the water table mimics the topography (elevation), the absolute water table (relative to sea level) ranges from 0 to 600 m for example in Bhima, so that errors of several meters would appear very small visually. We modified the sentence (line 313-317) : "However, a comparison between observed and simulated water tables is not relevant and not sensitive to parameters, as the water table mimics the surface elevation, which extends to several orders of magnitude, as noted by Gleeson et al., 2021 and Reinecke et al., 2020. Indeed, observed water table varies from 430 to 1000 m in Bhima while observed water table depth varies from 1 to 20 m."

equation after l. 331: the summation index (typically i) is missing! It has to be WTD_obs,mean_i and WTD_sim,mean_i. and the it also should read i=1 below the Sigma.
AR : Thanks. Following your suggestion, we modified Equation 2 (line 320-321) :

$$C_{mean} = 100 \times \frac{1}{n} \times \sum_{i=1}^{n} \left| \frac{\overline{WTD}_{obs,i} - \overline{WTD}_{sim,i}}{\overline{WTD}_{obs,i}} \right|$$

l. 334: the first sentence is trivial, the second questionably – a discrepancy of just below 50% is not an indication of a good model; presumably, both NSE and KGE would show quite poor values
AR : We agree with this comment. The sentence is trivial as Cmean is a very simple index representing the "normalized mean water table depth difference" as mentioned line 330-331. Defining a value of 50% is subjective, so we remove the two sentences and consider the criterion is simple enough to be understood.

equation after l. 341: the summation index is missing! This time it is t, it should read WTF_obs,t etc.
AR : Thanks again for this remark. We hope Equation 3 is clear now (line 330-331) :

$$nRMSE = \frac{100}{std(WTF_{obs})} \sqrt{\frac{1}{n_{obs}} \times \sum_{t=1}^{n_{obs}} (WTF_{obs,t} - WTF_{sim,t})^2}$$

Figure 4c) and all with a time axis: delete "[monthly]" in the legend of the x axis. This is just calendar years.
AR : Ok, the new version includes this modification.

Figure 4b): it would be quite natural to include the r2 value for this scatterplot. Similar in Fig. 5b)
AR : The revised manuscript includes this modification. r^2 equals 0.67 and 0.0 for Seewinkel and Bhima, respectively. While r^2 reaches 0.98 when comparing absolute water table instead of water table depth.

Figure 5c): here, the temporal resolution seems to be 6 months, which is rather odd. The model operates at daily scales, if you don't want to show the resulting time series you might aggregate to monthly values, but not half-yearly.
AR : Yes, the temporal resolution is twice a year because of the availability of observed water table. More precisely, each year there is one observation before and after the monsoon. This is mentioned at the end of section 4.2 "Comparison between observed and simulated water table" (line 340 : "twice a year from 1997 to 2009"). Because the model operates at daily scale, we compared simulated data at the same day than observed data. To be clear we added one sentence to the caption of Figure 5c) (line 412-416):
"c) Comparison between observed and simulated water table time fluctuations averaged across all monitoring boreholes and expressed as anomalies (relative to the time average). Based on the daily simulation, simulated water table is compared at the same days than observed water table, twice a year, before and after the monsoon."

We also added "monthly" to the caption of Figure 4c) (line 401).

l. 408: the uninitiated reader might wonder about "KGE" which is nowhere explained. Write out Kling-Gupta Efficiency and provide the reference.
AR : The new version includes the following improvement (line 418) :
"The Kling-Gupta Efficiency (KGE) (Gupta et al., 2009) values are…"

l. 584: "hydrological modelers never … observed in river networks"? Doesn't make sense. Something went wrong here. Even replacing "modelers" by "models" doesn't help. What are you intending to say here?
AR : Thanks for reporting this mistake. This new sentence should be better (line 594- 595):
"First, large-scale hydrological modelers never have a map of the actual river network, including the smallest low-order and intermittent rivers. They usually generate a river network based on a digital elevation model…"

Second reviewer

Thank you for these comments. We answered to each comment below in italic and labelled AR (Authors Response). Corrections to the manuscript are written in blue.

Specific comments

1) The first sentence of the short summary is ambiguous. I'd suggest something along the lines of "We develop and test the first large-scale hydrological model at regional scale with a very high spatial resolution that includes a water management and groundwater flow model"
AR : See the next comment.

2) It would help if the short summary had a statement about what these results tell us about this model development. E.g. what can it do, that we could not before.
AR : Thanks, we hope the short-summary will be better following your suggestions. Based on the two first comments we propose a new short summary (<500 characters, including spaces):
"We develop and test the first large-scale hydrological model at regional scale with a very high spatial resolution that includes a water management and groundwater flow model. This study infers the impact of surface and groundwater-based irrigation on groundwater recharge and on evapotranspiration in both irrigated and non-irrigated areas. We argue that water table recorded in boreholes can be used as validation data if water management is well implemented and spatial resolution is ≤100m."

3) L30: (river incision) is a confusing statement here.
AR : Thank you for this remark, we acknowledge that the statement is confusing. In the new version, we modified the sentence by (line 33-34, abstract) :
"We found that grid resolution is the main factor that affects water table depth bias because it smooths river incision, while pumping affects time fluctuations."

4) L41: "[…] the last point is of interest for locally relevant applications, but it could be replaced by "representing small-scale processes driven by topography"." This is a very confusing statement that does not seem to reflect the complexity small scale processes in hydrology.
AR : This sentence is linked to the challenge mentioned just before: "improving spatial resolution". We assume that this part is clear : "The last point is of interest for locally relevant applications". The following part focuses on the importance of topography at small scale which does not include all the

complexity of hydrology as pointed by the reviewer. The topography is specific to this study. Accordingly, we modify the sentence by (line 42-44) :
"The last point can be replaced in part by 'representing water flows driven by small-scale topography', and is of interest for processes at large-scale and locally relevant applications".

5) Section 4.2. Cm of water table depth is a bit an awkward metric as this puts (potentially) a lot more weight on very shallow water table depths (it can go to near infinitely large contributions).
AR : We agree with this comment. Our objective was to divide the error by the observed mean water table depth to give more weight to small errors when water table is shallow (thus, an error of 0.5m at 2m-depth is similar to an error of 2.5 m at 10 m-depth). Here, mean observed water table depth ranges from 0.25 to 12 m and from 1 to 20 m for Seewinkel and Bhima, respectively. Thus, shallow water table depths do not have too much weight. We completed the description of Equation 2 (C_{mean}) just after (line 323-324):
"Note that $\overline{WTD}_{obs,i}$ ranges from 0.25 to 12 m and from 1 to 20 m for Seewinkel and Bhima, respectively, so, shallow water table depths do not have too much weight on C_{mean}."

6) Line 463-465: these units do not seem to make sense.
AR : Thanks for reporting it. In fact, we used "mm/yr/m²" because this is an evapotranspiration rate [mm/yr] per m² of a specific land cover while usually in the study "mm/yr" refers to the basin-scale flow rate. To clarify, these units are replaced in the revised manuscript by: "mm/yr per unit area of groundwater-supported areas" (line 473 and line 475).

7) Specify somewhere what KGE is (line 408).
AR : Thanks, we forgot this point. The new version includes the following improvement (line 418) :
"The Kling-Gupta Efficiency (KGE) (Gupta et al., 2009) values are…"

8) Figure 4-5: the multiplication factor of the panel (a) y-axis is not readable).
AR : We agree that the multiplication factor was to small, thanks. The value of the multiplication factor is 1e6 and is the text is bigger in the new version.

9) Is the appendix necessary or could it just be SI?
AR : We put Appendices 1 and 3 in SI. We think that the two remaining appendices are necessary.