Telteu et al., 2021, under review GMD: Understanding each other's models: an introduction and a standard representation of 16 global water models to support improvement, intercomparison, and communication
https://gmd.copernicus.org/preprints/gmd-2020-367/

Thank you very much for your comments and recommendations.
The original comments of reviewers are in black color and indicated by “RC”.
Replies by the authors are indicated by “AC” and colored in blue.
Please see below our answers to your comments and recommendations.

Answers for Reviewer #1: Wouter Knoben
RC: “The target audience includes students, junior and senior scientists, modellers, interested stakeholders, and members of the general public interested in understanding global water models and modelling the impact of climate change on freshwater systems.”
I appreciate that the paper wants to reach a very wide audience but I wonder to what extent it actually does that. Would this paper be easily understandable by someone who does not have an existing background in at least one of the modeling communities? Would such a person care very much that these models use slightly different equations for certain processes?
In my opinion this sentence over-promises and in these times where communication of science to wider audiences is receiving considerable attention, I would recommend to stick with a narrower scope. This paper is an interesting resource for modelers and students looking to become modelers. I don't particularly see this as an effective tool to communicate between-model differences to stakeholders or the general public.
AC: Thank you for your comment. We consider that this study facilitates an understanding of the global water models and of the modelling the impact of climate change on freshwater systems through the simple language style used, the definitions presented, statements made. We agree that the stakeholders and the general public might be a little bit interested in the equations used for certain hydrological processes, but they might be interested in our analysis and synthesis on why models and their results are different. Additionally, many stakeholders and members of the general public collaborate with modellers to assess the impact of climate change on freshwater resources and share their experiences and knowledge. One example of collaboration between stakeholders and modellers could be ISIMIP. Together they co-create new knowledge. Therefore, in these particular cases, this study is useful to stakeholders and members of the general public because it offers the basis for an efficient collaboration, scientific definitions, explanations, statements on water modelling.
Our revised statement reads:
The target audience includes in particular students, junior and senior scientists, and modellers, or people who want to become modellers. Furthermore, this study could be used by stakeholders or other people who want to understand the background of global water models and how they simulate the global freshwater system.

RC: “The global hydrological community focuses primarily on surface water and groundwater availability, its human interference, and their daily to century-scale changes.”
I highlighted this in the previous review, to ask if this is typical because hydrologic processes have important sub-daily variation. The authors responded that "for most GHMs, daily time steps are still the state-of-the-art, whereas for LSMs, sub-daily timesteps are the standard. [...]".
I am inclined to disagree but I think this can be traced back to a difference between what I and the authors see as typical models in the GHM, LSM and DGVM communities.

I will therefore repeat a suggestion from my first review: I think it would be extremely helpful if the authors not only introduce the three communities by name at the start of section 2.1, but also list what they consider typical models in each community. Because the dividing lines between these communities are quite vague, this is not obvious. Presenting these examples in section 3 (as per the authors response to my original comment) does not help the reader understand section 2.1.
AC: Thank you for your comment. We moved the examples of GWMs from the beginning of section 3 to subsection 2.1. We feel that we already described the typical models in each community quite thoroughly.
Our revised statement reads:
In this study, land surface models are CLM4.5, CLM5.0, DBH, JULES-W1, MATSIRO, and ORCHIDEE. Global hydrologic models are CWatM, H08, Mac-PD20, mHM, MPI-HM, PCR-GLOBWB, VIC, WaterGAP2, and WAYS. One model (LPJmL) is a dynamic global vegetation model.

RC: "Parameters and coefficients represent numbers that describe a particular characteristic of reality, of the model, of the catchment area or flow domain. Some examples are **runoff coefficient**, soil porosity, hydraulic conductivity of different soil horizons, maximum soil water storage, maximum canopy water storage, mean residence time in the saturated zone, surface roughness, and vegetation properties (Beven, 2012)."
"runoff coefficient" - This seems an odd one in the list. All other examples are physical properties of the system that can be measured. The runoff coefficient is a statistical descriptor of a flow regime/catchment behaviour (sometimes referred to as a "signature"). I'd suggest to remove this.
AC: Thank you for your comment. We removed the "runoff coefficient" from the text.

RC: "mean residence time in the saturated zone". This is not a physical property either, but emergent behaviour that results from soil porosity, hydraulic conductivity, etc.
AC: We consider that the mean residence time in the saturated zone is the average time water spends in a subsurface system before it emerges as surface flow. We agree that this represents behaviour that results from soil porosity, hydraulic conductivity. Therefore, it represents a particular characteristic of the flow domain defined as the volume in which the flow takes place. We consider keeping this in our statement, as this list is not intended to name physical properties only but it also describes parameters used by the models.

RC: "Mathematical constants cannot be measured, but can be calculated and have a fixed numerical value, for example, $e = 2.718\ldots$, $\pi = 3.142$, $i^2 = -1$."
I agree that this is a mathematical constant but it seems an odd example in a paper about hydrologic models. I'd recommend to stick with examples that are actually used by the 16 models.
AC: Thank you. We removed "i^2" because it is not used by the GWMs. For example, "π" is used in Table S12 by MPI-HM. "e" is used in Table S20 by H08.

RC: I'd say the key characteristic of model output variables is that these are the parts of the simulation we want as output, much more so than that they tend to change in time and space. This may be more accurately phrased as "Output variables are the simulations of interest, for example, streamflow in a river catchment."
AC: Thanks for highlighting where the focus of this sentence should be. Our revised statement reads: *Ultimately, output variables are the results of the simulation and vary in space and time, for example, streamflow in a river catchment.*

RC: “During simulations, many parameters receive specific values because they cannot be measured everywhere, therefore, they are calibrated or tuned to attain the best match between simulated and observed data.” or estimated from auxiliary data that can be measured. An example would be the use of lookup tables for vegetation (approximate) properties based on satellite observations of land cover.
AC: Thank you for the good idea. Our revised statement reads: *During simulations, many parameters receive specific values because they cannot be measured everywhere, therefore, they are calibrated or tuned or estimated from auxiliary data that can be measured (such as lookup tables for vegetation properties based on remote sensing observations), to attain the best match between simulated and observed data.*

RC: “The final steps of a simulation are **to validate simulated and observed data**, to find out how well they fit, and to evaluate the simulated results through analysis and visualization.”
“**to validate simulated and observed data**“ - I don't fully understand this part of the sentence. This seems to say that simulations and observations are validated independently. In the case of observations, I assume this means checking whether the observations are not faulty? For simulations, does this mean checking that the model code does not contain errors? Oreskes et al (1994; 10.1126/science.263.5147.641) may be helpful since it provides clear definitions of terms like verification, validation, etc.
AC: Thank you for pointing this out. Our revised statement reads: *The final step of a simulation is to validate (evaluate) simulated model output with observed data through analysis and visualization.*

RC: Section 3: “Key characteristics of 16 global water models included in the study”
Given that this study focuses on differences between model structures without looking at any simulations, is all of section 3 necessary?
This may appear in a condensed form as the introduction to the current section 5 (where model differences are discussed) to reduce manuscript length. The section on calibration can probably be removed completely without losing any info relevant to the rest of the manuscript.
AC: Thank you for your comment. In this study, we present the key characteristics of 16 GWMs used in ISIMIP2b. Some GWMs have the ability to operate in different conditions, for example, at various spatial–temporal scales: CWatM, CLM4.5, CLM5.0 (3 h time step at around 11 km). Therefore, we feel that it is very important to present their key characteristics for the ISIMIP2b simulations. We intend to keep the information on calibration, because it is an often requested feature in global water modelling, see e.g. Krysanova et al., 2018, 2020.

RC: “Generally, these models are suitable for application over a catchment size of not smaller than 9,000 km^2 or at least four grid cells (Döll et al., 2003; Hunger and Döll, 2008).
Does this statement refer to a particular configuration of these models? There are plenty publications that run mHM or VIC on catchments much smaller than 9000 km^2. I assume this goes for the other models too.
AC: In this study, we analyzed the GWMs used for ISIMIP2b. In this project, GWMs use a spatial resolution of 0.5° × 0.5° and carry out simulations on the global scale. This is specific for ISIMIP2b and some GWMs. When we discuss this spatial resolution and this type of GWMs, we consider that these models obtain good model results if they are applied over a catchment size of not smaller than 9,000 km^2 or at least four grid cells because of the input data quality and process representation. We provide further references with the experiment done on this issue: Döll et al., 2003; Hunger and Döll, 2008.

Indeed, it is possible to run the aforementioned models for smaller basins, however, this should be justified by the availability of the meteorological forcing data. Indeed, Mizukami et al., 2019 and Rakovec et al., 2019 used mHM and VIC for smaller basins than 9.000 km^2 at finer spatial resolutions, which was possible in their continental-wide setups due to forcing data available at finer spatial resolution (~0.1deg).
Furthermore, we are now highlighting that our statement is only true for the given spatial resolution of ISIMIP2b thus the sentence now reads as: *These models, as applied within the ISIMIP2b framework, are suitable for application over a catchment size of at least four grid cells (Döll et al., 2003; Hunger and Döll, 2008).*

RC: "least four"
AC: Thanks. We revised the sentence.

RC: Also, perhaps this mention of grid cells can be removed because whether it is four cells or not depends entirely on the grid in question. I doubt this adds much over the 9000 km^2 mention.
AC: Thanks. We deleted this information.

RC: “These models simulate the terrestrial water cycle, on the global land area (except Antarctica) with a spatial resolution of 0.5° × 0.5° (~55 km × 55 km at the Equator), and quantify water flows, water storage compartments, and human water use under the given climatic and socioeconomic conditions.”
Suggest to remove this part, because it is covered in more detail in 3.2.
AC: Thanks. We deleted this part.

RC: “ORCHIDEE runs with a spatial resolution of 1.0° and **has its outputs converted to 0.5° spatial resolution**.” How? Is each grid cell simply chopped into 4 pieces or is some form of interpolation used?
AC: The ORCHIDEE simulation results were remapped conservatively to a spatial resolution of 0.5° × 0.5°, for consistency with other models participating in ISIMIP2b, by simply reducing the grid cells size (four 0.5° × 0.5° grid cells will have the same value as 1°×1°).

RC: Subsection “3.4 Calibration approaches”
What seems to be missing is the outcomes of calibration. How accurate are these models with their calibrated simulations? Will differences in accuracy be related to differences in model structure later on in the paper?
AC: Thank you for your comment. Yes, GWMs intend to have better results if they are calibrated (e.g., WaterGAP2: Müller Schmied et al., 2014, CWatM: Burek et al., 2020). In this study, only three models are calibrated (Table 6). Moreover, details on the calibration and the simulation performance are available in individual model development studies for each model. These studies are mentioned in Table 11.
There is rising interest in model calibration on that scale (e.g. see the recommendations of Krysanova et al.,2018, 2020) and we want to reflect on these activities briefly. However, we consider that evaluating the accuracy of the models or relating the accuracy to the differences in model structure is beyond the scope of this study and might be something for a follow-up-study that incorporates a model output directly.

RC: “CWatM **calibrates monthly or daily streamflow** for 12 catchments using the Distributed Evolutionary Algorithms in Python (DEAP) approach (Burek et al., 2020)”
Does it do both or just one of them in this MIP?

AC: Thank you. We deleted the sentence because CWatM was not calibrated for the ISIMIP2b.

RC: “**Section 4. Creating the standard writing style of model equations”**
It takes until section 4 (200 lines, 6 pages) to reach the first description of the standard writing style mentioned in the title. Repeating a statement from the first review, this paper tries to do a lot. I would strongly recommend to either change the title to more accurately reflect the paper contents, or to try and trim down the paper as much as possible, or to try and manage reader expectations more.
Perhaps narrowing the scope of the paper (e.g. limit it to existing and prospective scientists in one of the three fields) gives an opportunity to slim down some of these sections by simply providing summary sentences and references instead of the detailed examples writing for a much wider audience may require.
AC: Thank you for your comment. We have modified the title to relate to the reader’s expectations and have revised the target audience: *Understanding each other’s models: an introduction and a standard representation of 16 global water models to support intercomparison, improvement, and communication*
We consider that the presented differences in modelling approaches, the definitions, and the key characteristics of the analyzed GWMs are needed for a better understanding of the similarities and differences among the 16 GWMs described in section 5, independently of the audience. Similarities and differences among the analyzed 16 GWMs exist because of their research purpose, different modelling approaches, and decisions made in ISIMIP2b. We envision this paper to be a resource with information about these models, compiled in one place, without having to refer the reader to another paper all the time.

RC: “the”
AC: Thank you. We corrected the sentence. Our revised statement reads: *In this study, the rationale in finding similarities and differences among 16 GWMs is based on how models simulate the terrestrial water cycle.*

RC: “Generally, the models have different style in describing their structure, defining their variables, and writing their equations. Furthermore, a unique equation can be implemented in various ways (e.g., discrete vs. analytical form, focusing on flows or water compartments) or **parameterized** differently.”
Within my understanding/use of terminology, a parametrization is a unique equation that describes some process. This sentence implies that for a given unique equation, multiple parametrizations can exist. Can it be clarified how "parametrization" is used in this paper? Maybe in the sense of "uses different parameter values"?
AC: Thank you. Our revised statement reads: *Generally, the models have different style in describing their structure, defining their variables, and writing their equations. Furthermore, a unique equation can be implemented in various ways (e.g., discrete vs. analytical form, focusing on flows or water compartments) or can use different model parameter values.*

RC: “and”
AC: Thank you. We corrected the sentence. Our revised statement reads: *We decided to describe 16 GWMs based on the equations implemented for eight water storage compartments and six human water use sectors.*

RC: “We define parameterization as changes of model parameter values (Samaniego et al., 2010).” I see. This should probably be defined before parametrization is discussed in the text to avoid confusion for readers like myself.

AC: We moved our definition to subsection 4.3.

RC: “However, the purpose of this baseflow storage, for MPIHM, is predominantly to cause a delay in river discharge and not to simulate groundwater in detail.”
This is intriguing. How did the authors reach this conclusion? Does this come from the MPI-HM documentation, is the MPI-HM store functionally different from e.g. the lowest soil layer in VIC, something else?
AC: The baseflow storage was originally part of the river routing scheme in which three reservoirs are used to represent lateral processes with different time scales: the overland flow, the baseflow and the river flow reservoir (see Hagemann and Duemenil, 1997). It is functionally different from a soil layer storage as it only receives water draining from the soil, but cannot provide water to the soil via diffusion nor is it subject to any soil properties. Nonetheless, after several discussions among modelling teams, we concluded that this baseflow storage mimics part of the functionality of a groundwater storage and included it as such for better comparison with the other models.

RC: I really like this new and improved section on definitions. It could make for a great starter on further discussion about terminology used in models and how this affects how we read each other's papers where such terms are typically not cleanly defined.
AC: Thank you very much for your encouraging comment.

RC: "storage" ? In my understanding "stock" typically refers to goods kept ready to be sold.
AC: We replaced *stock* with *storage*.

RC: It would be good to clarify the definitions of "runoff" and "streamflow" in the previous section too.
AC: Thank you. We added new statements to subsection 4.3: *In this study, we define streamflow as the volumetric flow rate of water through a river cross-section. Therefore, the streamflow is the water transfer which is routed through a channel towards the ocean or towards an inland sink. We define the total runoff as the (not routed) total amount of water that runs-off the grid-cell, either over the soil surface, or from the subsurface (lateral flow). In some studies, the streamflow is converted to runoff by dividing the streamflow values with the area upstream of the gauging station (for example, the area upstream of station according to the DDM30‘ river network Döll and Lehner, 2002).*

RC: “Therefore, the modelling teams checked the model equations on their plausibility.” suggest to replace this by "correctness" or something along those lines. "Plausibility" to me suggests that the teams are only partly certain they got it rig.
AC: Thank you. We revised our statement.

RC: “In the final step, we reevaluated the collected and homogenized model equations for plausibility.”
Plausibility of what? Of how accurate those equations describe hydrologic reality, or how accurate the information in the appendices of this paper describes what's actually encoded in the model code?
AC: Thank you very much. It refers to how accurate the information in the appendices of this paper describes what is actually encoded in the model code. The modelling teams checked the consistency of the model equations from the supplementary information with the model code and found similarities among the analyzed models.
Our revised statement reads: *In the final step, we reevaluated the collected and homogenized model equations for their consistency with the model code.*

RC: “5.1 Similarities and differences in simulating eight water storage compartments”
As a general comment, it would be good in the entire section 5 to be specific about which models explicitly simulate an energy balance and which models use some form of potential evaporation estimate to approximate energy terms. Details about how PET is estimated by different models would be good to mention too.
AC: We added new statements, please see our answer below.

RC: “Ten models compute canopy water storage by subtracting the throughfall amount and canopy evaporation from the total precipitation.”
[Example of where details about energy balance/PET would be useful] How is canopy evaporation estimated by these models? I assume with some (adjusted) form of potential evaporation estimate?
AC: Our statements are: *Three models do not compute potential evapotranspiration (Table S2, Tables 7 and 8). Seven models apply the Penman–Monteith method to compute potential evapotranspiration (PET). PCR-GLOBWB applies the Hamon method to simulate PET, while mHM applies the Hargreaves-Samani method. ORCHIDEE applies a simplified Penman & Monteith equation (Monteith, 1965) with a correction term developed by Chris Milly (1992). WaterGAP2 and LPJmL apply the Priestley-Taylor equation, while H08 and MATSIRO apply the Bulk method.*

RC: “Generally, prescribed vegetation ignores the decisive interaction between vegetation and runoff and interactions between the atmosphere and Earth’s surface, partly presented in **section 3.2** (Gerten et al., 2004; McPherson, 2007; Nicholson, 2000).”
I'm unsure how to interpret the word "prescribed" here. Suggest to clarify in text what is meant.
AC: Our new statements are: *In the ISIMIP2b, the word “prescribed” has two meanings: (i) data which are simulated by other models and provided by the ISIMIP2b framework as input (https://www.isimip.org/gettingstarted/details/38/); (ii) data obtained from satellite observations, other datasets, or maps. Prescribed data highlight some limitations of the models or underline the lack of some processes that were intentionally or non-intentional removed from the model structure, according to the purpose of the model development or other priorities such as time.*
RC: This is not correct; 3.2 specifies spatial/temporal model resolution.
AC: Thank you. We deleted this information.

RC: “3. Ratio between rainfall or snowfall and total precipitation (CLM4.5, CLM5.0, MATSIRO);”
This seems odd to me. Does this mean that these models intercept all rain but no snow (or vice versa)?
AC: This means that these models distinguish between rainfall and snowfall. We added this statement to subsubsection 5.1.1.

RC: ”Generally, it was found that simulations depend on the number of plant functional types (PFTs) prescribed or defined in the model and on the processes used to estimate plants’ ability to adapt, acclimate, and grow in new environmental conditions (Sitch et al., 2008).”
Does this sentence refer to a finding from the authors' study or does this introduce a known finding that comes from Sitch et al.?
Also "simulations" is quite vague. Simulations of what?
AC: Our revised statement reads: *Sitch et al., 2008 found that simulations on CO_2 fertilization effect depend on the number of plant functional types (PFTs) prescribed or defined in the*

model and on the processes used to estimate plants' ability to adapt, acclimate, and grow in new environmental conditions.

RC: Related to snow processes, (how) do these models simulate ice and/or frozen soils?
AC: Four models simulate the frozen soil as described in Table S13. We consider keeping this information in the supplement to streamline the manuscript, as only a few models simulate frozen soil.

RC: "Overall, 10 models consider initial infiltration as inflow of the soil storage, while three models (H08, JULES-W1, and WAYS) consider throughfall (Table S14)."
Is this a difference in terminology only, or do the infiltration-models do something extra (like infiltration excess runoff)?
AC: GWMs compute differently throughfall, infiltration, and soil storage. We mention Tables S5, S14, and S25 for other details. Some models compute infiltration excess runoff and other models compute saturation excess overland flow (see Tables 9 and 10). We explain how throughfall is computed by these models in Lines 363 – 369.

RC: "JULES-W1 also uses a "zero-layer" scheme that does not use explicit model layers to represent snow, instead adapting the topsoil level to represent existent snow processes. In the original "zero-layer", snow scheme has a constant thermal conductivity and density. Bulk thermal conductivity of snow on the surface layer decreases due to both the increased layer thickness and the different conductivities of snow and soil. Surface energy balance and heat flux between the surface layer are controlled by insulation factors and layer thickness (Best et al., 2011)."
Would this be more appropriate in the snow section?
AC: We consider keeping this information in this section because it is related to both snow and soil processes. It also fits better with our explanations on other models.

RC: "the"
AC: Thank you. We corrected the sentence. Our revised statement reads: *In GWMs, the groundwater compartment simulates hydrologically the saturated zone or phreatic zone (WaterGAP2) or an unconfined aquifer (CLM4.5).*

RC: "In ISIMIP2b, two models (JULES-W1 and LPJmL) consider the water excess from the bottom soil layer as seepage and **relate** this variable with groundwater recharge because they do not have a groundwater compartment
Is this the right word here? Should this be "equate"?
AC: Thank you. We replaced *relate* with *equate*.

RC: "to " The mHM model uses a mesoscale routing model with an adaptive time step according ~~with~~ the spatially varying celerity (Thober et al., 2019).
AC: Thank you. We corrected the sentence. Our revised statement reads: *The mHM model uses a mesoscale routing model with an adaptive time step according to the spatially varying celerity (Thober et al., 2019).*

RC: "LPJmL used prescribed data for domestic and industrial water consumption data and assumed that only the consumed water amount is withdrawn." It's not quite clear to me what this means.
AC: Thank you. We corrected the sentence: *LPJmL used input data for domestic and industrial water consumption data, provided by the ISIMIP2b framework, and assumed that only the consumed water amount is withdrawn.*

RC: “CWatM calculates the water withdrawal in total from all users and afterwards it distributes the total withdrawal to different sources: surface water, sustainable groundwater (available groundwater = long-term groundwater recharge of the last 30 years in the analyzed time interval), **unsustainable groundwater human water use sectors** (domestic, livestock, irrigation, industry)”
Are some words missing in this part of the sentence?
AC: Thank you. Our revised statement reads: *CWatM sums up the water withdrawal from all users and distributes the total withdrawal to different sources: (i) surface water, (ii) sustainable groundwater (renewable groundwater = long-term groundwater recharge of the last 30 years in the analyzed time interval), and (iii) unsustainable groundwater (nonrenewable groundwater = additional water gained by groundwater abstraction in surplus of groundwater recharge; Wada et al., 2012).*

RC: “We consider that two issues are useful to interpret model results, first, knowing model structures, and, second, identifying the effect of model structures on model results. However, the present study is focused only on the first issue, respectively, knowing model configurations needed to interpret various model results.”
This seems like something that should be mentioned in the introduction where the scope of the paper is defined; not 19 pages into the manuscript.
AC: Thank you. We moved the paragraph to the introduction.

RC: “In global water modelling, there are some more methodologies that can be tested to evaluate multi-model structures and model equations, also considered as hypotheses on runoff generation, for example, Rainfall-Runoff Modelling Toolbox (Wagener et al., 2001); the rejectionist framework (Vaché and McDonnell, 2006); Framework for Understanding Structural Errors (FUSE, Clark et al., 2008); SUPERFLEX (Fenicia et al., 2011); Catchment Modelling Framework (CMF, Kraft, 2012); Structure for Unifying Multiple Modeling Alternatives (SUMMA, Clark et al., 2015 a and b). Other methodologies can be used to evaluate parameter values such as Model Parameter Estimation Experiment (MOPEX: Duan et al., 2006), multiple-try DREAM(ZS) algorithm (Laloy and Vrugt, 2012), Generalized Likelihood Uncertainty Estimation methodology (GLUE: Beven and Binley, 2014), perturbed parameter ensembles (Gosling, 2013), the Uncertainty Quantification Python Laboratory platform (UQ-PyL: Wang et al., 2016), Multiscale Parameter Regionalization (MPR, Samaniego et al., 2010 and 2017). Thus, some existing methods might offer some solutions for reducing the high number of parameters and their values still found in global water models, and to apply more reasonable regionalization schemes in global water research (Bierkens, 2015). Other methods can be found in frameworks proposed by Döll and Romero-Lankao, 2017 and Kundzewicz et al., 2018.”
This paragraph still feels like a very specific snapshot of possible methods that exist to investigate model structures. In my opinion it lacks the depth needed to contribute to this manuscript and I would suggest to remove this paragraph entirely. The references in line 750 probably provide enough of a starting point for a reader interested in designing MIPs.
AC: We consider keeping this information because it underlines ways to improve the global water models and design future model experiments.

RC: “Furthermore, some studies have tested how model equations combined in different configurations and using different parameter values influence the simulations: Essery et al., 2013 (testing 1701 snow models); Niu et al., 2011 (Noah-MP model); Pomeroy et al., 2007 (Cold Regions Hydrologic Model, CRHM); Kuppel et al., 2018 (Ecohydrologic model, EcH2O). In summary, they found that some model configurations provide consistently good

results, others provide consistently poor results, and many configurations provide good results in some cases and poor results in others (Essery et al., 2013).
This seems more of a justification for doing MIPs and less of a recommendation for how to run one (as this section title indicates). Suggest to move to the manuscript introduction instead.
AC: Thank you. We moved this information to the introduction.

RC: "describe your model or a model through your eyes and other's people eyes;"
Given the amount of italics in this sentence, it must be important. Maybe this justifies further details? I imagine this is meant to prevent confusion caused by differences in terminology and (implied) assumptions in model code that may or may not be clearly documented. If so, I suggest to be specific about this.
AC: Thank you. Our revised statement reads: *describe your model or a model through your eyes and other's eyes to identify differences in terminology and assumptions in model code and similarities and similarities and differences among the models.*

RC: "Our future research will include describing the GWMs analyzed in this study, through a standard visualization of the water cycle that will show the water storage compartments, water flows, and human water use sectors included in the ISIMIP2b model structures."
Would this be different from the current Figure 1? If so, it may be helpful to outline how and possibly include this new visualization in this paper. It may make the section on model differences more intuitive.
AC: We consider that it is beyond the scope of the present study. We plan to describe 16 GWMs using one standard water cycle diagram that is currently under development and deserves its own story. We created Figures 1 and 2 for the present study.

RC: "Ultimately, GWMs may even become a part of the Earth System Models used to simulate the water cycle at a high resolution, including human water demand and use (Wood et al., 2011; Bierkens et al., 2015)."
Do ESMs not already include much of the hydrology discussed in this paper? This sentence seems to imply that ESMs contain only the most rudimentary hydrology which I don't believe is the case. If the authors wish to end on this note I recommend to be specific about which aspects of GWMs they would like to see become part of ESMs. For context, in my understanding SUMMA (l713) is an ESM and I don't think it would stand out as being much simpler than the models discussed in this paper.
AC: Yes, ESMs include hydrology discussed in this paper but they in general have a lower spatial resolution compared to the GWMs. For example, CLM, JULES, and ORCHIDEE represent the land surface component (hydrology) of an ESM and have almost the same processes, but the model versions analyzed in the present study differ from the original model version used for an ESM. When running the coupled ESM, the spatial resolution is much lower than offline simulations driven by downscaled climate forcings.
Our revised statement reads. *Ultimately, specific features of GWMs such as dam operation, human water abstractions, routing approaches, and calibration might become a part of future Earth System Models (Wood et al., 2011; Bierkens et al., 2015).*

References:

Hagemann, S. and Dümenil, L: A parametrization of the lateral waterflow for the global scale, Climate Dynamics volume 14, 17–31, https://doi.org/10.1007/s003820050205, 1997.

Krysanova, V., Hattermann, F., and Kundzewicz, Z.: How evaluation of hydrological models influences results of climate impact assessment – an editorial", Climatic Change, 163:1121–1141https://doi.org/10.1007/s10584-020-02927-8, 2020.
Krysanova, V., : How the performance of hydrological models relates to credibility of projections under climate change., Hydrol Sci J, 63(5), 696–720, https://doi.org/10.1080/02626667.2018.14462142018
Mizukami, N., Rakovec, O., Newman, A. J., Clark, M. P., Wood, A. W., Gupta, H. V., and Kumar, R.: On the choice of calibration metrics for "high-flow" estimation using hydrologic models, Hydrol. Earth Syst. Sci., 23, 2601–2614, https://doi.org/10.5194/hess-23-2601-2019, 2019.
Rakovec, O., Mizukami, N., Kumar, R., Newman, A., Thober, S., Wood, A. W., et al.: Diagnostic evaluation of large-domain hydrologic models calibrated across the contiguous United States. Journal of Geophysical Research: Atmospheres, 2019; 124: 13991– 14007. https://doi.org/10.1029/2019JD030767, 2019.
Wada, Y., van Beek, L.P.H., and Bierkens, M.F.P.: Nonsustainable groundwater sustaining irrigation: a global assessment. Water Resour Res 48:W00L06. doi:10.1029/2011WR010562. Special Issue: Toward Sustainable Groundwater in Agriculture, 2012.