AUTHOR'S RESPONSE TO RC2:

Manuscript hess-2018-626 by Martinez-de la Torre & Miguez-Macho: "**Groundwater influence on soil moisture memory and land–atmosphere interactions in the Iberian Peninsula**"

The LEAFHYDRO model was among the first ones to couple a physically-based 2D groundwater (GW) flow model with a land surface model (LEAF2). Transient applications so far focused on the USA (2007 papers), the Amazon (2012 papers), while the groundwater model has been coupled to Noah over South-America (Martinez et al 2016), and over New-Zealand (Westerhoff et al., 2018, not cited). The present paper reports the application of the LEAFHYDRO over the Iberian Peninsula (IB), with novel insights regarding the propagation of precipitation anomalies to water table depth (WTD) and soil moisture anomalies, at a pluri-annual timescale. This effect is all the more pronounced as the climate is more arid, as are the more "classical" impacts of GW on ET, soil moisture, and river discharge at seasonal or shorter timescales.

Authors: Thanks. We have included the Westerhoff (2018) reference in Section 2.4, when we describe the 2D groundwater model (uncoupled to LSMs). This work was finally published while finishing up our manuscript initial writing.

This makes this paper very commendable for HESS, but on the other hand, the paper lacks (i) good quality figures, (ii) sufficient information on the model and forcing datasets, (iii) solid quantification of the reported impacts and model validation, (iv) a real discussion of the results, including the limitations of the approach. Other problems include a structure that is not always very logic, and a tendency to overstatement. Apart from a few spelling errors, the language is very clear. **Eventually, I advise to substantially revise the paper before publication in HESS**. None of the suggested revisions is very complicated, but many of them are advisable, as detailed below.

Authors: Thanks for this complete assessment. We acknowledge the issues pointed out by the reviewer and have introduced substantial editions and changes to the manuscript to address them, making the paper stronger in our view. We discuss such changes in response to the reviewer's specific comments below.

1. **Figures:**
   (a) Most IB maps are too small, and it's almost impossible to see something. Please remove Southern France, Northern Africa and oceans, and magnify the remaining IB. When possible, please use the same color scale for comparable variables (e.g. equilibrium and non- equilibrium WTD). It would also be very informative to add the mean and std of each map over the IB.

Authors: Thanks. In the revised manuscript, we have cropped the figures including parts of Africa and France to make the results over the Iberian Peninsula more visible. The non-equilibrium color scale was chosen in order to contrast with the mean top-2m soil moisture plot in the same map. The means of the variables represented in each plot are included in the bars below on the anomaly figures (Figs. 10 and 11 in the original submission).

   (b) Fig. 3 is too small as well, and the color code of the points locating the points with observed does not seem well adapted, since a point can meet several conditions: for instance, what is the color of a point where bias is less than 2 m (red), and correlation with observed time series is more than 0.5 (green), which must be possible? There also seems to be some black points, in which case their meaning should be explained. But maybe they are purple… It would also be useful to report the classification used for Fig. 3 on the various panels of Fig 4 (insert for each point the bias, the correlation coefficient, and the wtd slope).

Authors: Thanks. In the revised manuscript, we have cropped figures 3 (now 4) and made it larger. The colour of the dots is easier to identify now, and we have also explained in the caption the meaning of red dots (no validation criteria matched) and the hierarchy of the criteria (in the case pointed out by the reviewer the dots appear as red, this happens in most red dots on the next figure showing time series and it is also explained now in the text when describing the figure). New caption: "*Shallow water table zones (light blue shades) and Iberian Peninsula wtd observation stations (dots). Red dots are locations where observed and simulated wtd differences are within 2 m; green dots are stations with correlation over 0.5 between observed and simulated wtd series; purple dots are stations with steep wtd slope (>=0.035 m month-1), well captured by the model; orange dots are cells containing more than one observation station; black dots are cells where none of the above criteria is met by the model. Over cells where more than one validation criteria is reached the point adopts the colour of the first criterium met (in the order presented here); for instance, cells with mean wtd differences lower than 2 m and also correlations above 0.5, are shown as red on the map*"

(c) For Fig3, Fig 5, Fig 15 (and potentially Fig4), it must be clarified if the reported correlation coefficient is calculated on the full time series (120 monthly values), or on the mean seasonal cycle (12 monthly mean values).

Authors: The correlations used in the wtd validation (Fig. 3, now 4) were calculated with the full time series available, where the time scale varied amongst the stations (see Fig. 4, now 5). The correlations reported in the river flow time series are calculated on the mean seasonal cycle, and that was the reason to show them inside the seasonal plots. We have clarified both scales in the revised manuscript and figure captions. Thanks

(d) Fig. 6: R seems negative if downward, which seems odd for the flux which recharges the GW.

Authors: Yes, that is true. We acknowledge the groundwater recharge is often referred to as the positive flux into the groundwater reservoir. In this work, we have followed the model criteria for signs in fluxes, so that upward is positive (like evapotranspiration from the surface) and downward is negative (like the water flux through the soil layers and then into the groundwater). We have changed the name of the flux to "net recharge" in the revised manuscript in order to clarify this point at different instances. The first time the net recharged is referred to in the manuscript in Section 2.1: "*The water flux through the water table or net recharge R is the sum of gravitational downward groundwater recharge and capillary flux, and depending on soil wetness and atmospheric demand, it can be downwards, causing the water table to rise, or upwards, causing the water table to deepen*"

(e) Fig. 7c: why not show a real mean seasonal cycle, with 12 monthly mean values, instead of 4 seasonal mean values? And couldn't you plot the seasonal cycle of the shallow WTD as well?

Authors: We think that the plot focusing on the 4 seasonal means is stronger, illustrating the greater groundwater influence in water-scarce seasons. Also, the presentation as 4 seasons follows the net recharge figures (Fig. 6, 7 now), connecting with the seasonal variability in the net recharge and making the point clearer for the reader.

(f) Fig. 8: the color scale is not clear, we cannot distinguish the values that are not zero. Besides, could you add a scatter plot of summer ET difference against summer WTD, to show if there is a kind a threshold WTD inducing a marked ET difference?

Authors: We have cropped and made the figure larger in the revised manuscript, making the green scale more distinguishable for the reader.

(g) Fig. 9 is very noisy: could you add the difference between center and left panels?

Authors: We have noticed a mistake in the caption, the left and centre panels descriptions were switched. We have corrected it in the revised manuscript. Of course we could add the difference plot suggested by the reviewer, but would not this make the figure noisier? The idea is that the plot on the right highlights with high values those areas of difference between the centre and left plots.

(h) Fig. 11: please add the lon/lat of the mapped area, either on the maps, or on the caption.

Authors: Yes, done. Thank you.

(i) Consider merging top panel of Fig 5 and Fig12; same for Fig 5 and Fig 15.

Authors: That was our initial approach in the first draft, but we believe after consideration that having the 2 figures makes it easier for the reader, as we present Fig 15 (now 16) when we study the main basins, while the plot on top of Fig. 5 (now 6) is about rivers and gauging stations.

By the way, why not show the FD simulations at all stations? And correct the statement that Ebro at Tortosa is where the model exhibits the best scores (L32-33 p 14): based on correlation coefficient, this station is only the third best for simulation WT based on Fig 5; besides, the ms discusses two models, so clarification is needed. Finally, the correlation coefficient is far from enough to support a performance analysis, and I strongly recommend that other classical criteria are documented (bias, important since ET changes between the two simulations; RMSE; Nash and/or KGE, which are classical skill criteria in river hydrology).

Authors: There are two important issues related to streamflow in these simulations. They are discussed in section 3.2 in the paper, but perhaps they need further clarification. The first one is related to the precipitation forcing data. From figure 5 it is obvious that there is a large amount of missing water in the model results. Only basin 5 (Guadalquivir) shows less streamflow in the observations than in the model, but this is because in this strongly regulated basin, water is heavily used for irrigation. While there can be some errors due to evaporation biases, we have evidence from local independent observation networks that this missing water is more related to the precipitation forcing (please, see the discussion about the same bias in the IB02 dataset in Rios-Entenza and Miguez-Macho, 2014). In the mountains, especially in the north, the IB02 dataset does not properly capture orographic enhancement, since it was obtained using simple interpolation algorithms.

The second problem is due to model parameterizations and is also commented in the paper. Surface runoff from excess saturation in thin soil or in subgrid near saturated areas is unrepresented. Due to unresolving hillslope hydrologic gradients at the 2.5km resolution, the connection between rivers and groundwater in cells where the mean water table is deep does not produce a good result either.

Since both forcing and model problems affect mostly mountain areas where terrain is complex, we are confident that the main conclusions in our work about groundwater and soil moisture are sound. However, we cannot say the same about riverflow, since the contribution from the mountainous areas to their total water budget is very important.

We have now calculated other skill scores for both experiments, as suggested by the reviewer. In the FD simulation, the lack of surface runoff is compensated by the fact that infiltration is readily incorporated into the rivers. Because precipitation amounts are biased low, winter peaks may look better in this FD simulation and some skill scores are better than in the WT simulation, but this does not mean that the result is physically correct.

Station - Basin	Basin Catchment Area (km²)	E WT	r WT	r_{mm} WT	E FD	r FD	r_{mm} FD
Foz de Mouro - MIÑO	15.407	-0.13	0.89	0.98	0.55	0.93	0.95
Puentepino - DUERO	63.160	-0.52	0.73	0.96	0.18	0.72	0.69
Almourol - TAJO	67.482	0.28	0.91	0.93	0.80	0.91	0.89
Pulo do Lobo - GUADIANA	61.885	0.07	0.65	0.71	0.21	0.54	0.66
Cantillana - GUADALQUIVIR	44.871	0.44	0.76	0.66	0.15	0.56	0.67
Tortosa - EBRO	84.230	0.36	0.74	0.93	0.55	0.82	0.87

For all the aforementioned reasons, we purposely wanted to limit our discussion about streamflow in the paper and just show the WT results. The only point that we wanted to make with the FD simulation is that in the Mediterranean climate of the Iberian Peninsula, summer stream flow is sustained by groundwater and, without it, in a simulation with a free drain approach, rivers dry out. We are confident that this result holds true, despite all the problems in the forcing and model parameterizations. We show the Ebro basin to illustrate this point because it is the one showing less streamflow total annual underestimation and annual cycle better matching observations in the control WT run, especially in winter. It so happens that it is also the largest basin the Peninsula, so the example is significant.

(j) Fig. 13: a full paragraph is devoted to analyzing the seasonal variations of the different variables (L21-31 p 13), but we cannot see them. Please add the mean seasonal cycle next to the 10-yr time series to support this discussion. It would also help to magnify the scale of SM differences. The caption says the precipitation anomalies are calculated over the entire basins, while the other anomalies are calculated in the fraction with shallow WT (ca 1/3): why not calculate them over the same domain, to avoid any doubts. Finally, the text says the WTD of Ebro and Segura recover after the pluri-annual central drought, but it is not discernible in the panels.

Authors: We chose to plot only shallow wtd regions within the basins to highlight the effects of the interaction with the water on soil moisture and ET; however, seasonal precipitation is computed over the whole basin because lateral groundwater flow and river infiltration redistribute infiltration horizontally, making precipitation over all cells in the basin potentially relevant for the results over shallow water table cells. The recovery over the last 3 years is, we believe, clear in most all basins that suffered the central drought; however, the reviewer is right in that in the Ebro and Segura basins only a change in tendency from deepening to stabilizing or slightly rising is discernible. We have clarified this in the revised version.

2. Methods:

The LEAFHYDRO model has already been published, but a paper needs to be self-consistent, and more info is needed on the parts that are relevant to the conclusions.

<u>The recharge calculation</u>, in particular, is far from being clear, at least to me, although I have looked for more information in Miguez-Macho et al. 2007, Part 2. This should be clarified in the article, and the following questions might help the authors:

(a) The calculation is different depending if the WTD is larger or not than the soil depth, but I couldn't understand scenario b with larger WTD. In this case, how are the water content of points B and C estimated? It's written the one of point C "is determined by mass balance from the fluxes above and below" (L5-6 p 5) but these fluxes also need to be estimated, and there seem to be too many unknowns: please clarify the system, including flux equations, boundary conditions, or any assumption regarding water content profiles, etc.

(b) In both cases, the water content of the unsaturated zone and WTD must be coupled, so what is the effective sequence of calculations over time? I struggle with "R is the amount of water from or to the unsaturated portion of layer 1 necessary to cause the rise or fall of the water table from its former position", knowing that WTD is updated based on equation 1 which depends on R.

Authors: [(a) and (b)] Yes, we agree that, since we initially tried to avoid a methodology section too long and filled with equations already published in Miguez-Macho et al (2007), the methodology section is not as explanatory as it should. The reviewer is right to point out that some clarifications are needed. We have changed Section 2.1 and we believe that the issues raised by this and other reviewers about the model formulation and steps have been addressed in the revised manuscript.

(c) Since R is calculated differently if the WTD is larger or smaller than 4m, can we see a discontinuity of net recharge values at 4m (plotting R as a function of WTD)?

Authors: Possibly, but the model formulation is designed precisely to avoid any discontinuity in water table depth or recharge. When the water table goes below 4m, calculations are identical to when it was in the layer above, and only as it goes below 4.5m they start to differ, but they do it very gradually. No discontinuity is observed in water table depth as it goes deeper (or shallower) than 4 or 4.5 m, which is a good indication that there isn't one in recharge either.

(d) This flux R is defined as the result of downward gravitational flux and capillary flux, which can be either up or downward. The resulting flux is called recharge in the results, but "flux through the water table in section 2.1 (L22 p 14): I invite the authors to harmonize throughout the ms, and use recharge, but as mentioned in my comment 1d, this "net" recharge, which can be positive or negative, should positive when down, to match the meaning of GW recharge.

Authors: We have followed the reviewer's suggestion and called the flux "net recharge" throughout the text, presenting it initially in Section 2.1 as "*The water flux through the water table or net recharge R*". We would prefer though to keep the signs as they are, positive upwards and negative downwards, as these are the signs in the model calculations too, where upward fluxes like ET are positive and downward fluxes like infiltration are negative.

(e) As an interested reader, I would also appreciate some explanations regarding the links between R and evapotranspiration, which must be tightly coupled as well: how is transpiration described? How is rooting depth described? What is the vegetation description at the surface: PFTs, mosaic approach, constant or varying over time, which input datasets?

Authors: The model parameterizes the calculation of transpiration and evaporation from canopy interception using PFTs and the vegetation data described in Section 2.2. We have added a paragraph at the end of Section 2.1 pointing out this: "*When there is vegetation on the surface, the water and heat exchanges between vegetation and the surrounding canopy air parameterization is based on Avissar et al. (1985). This methodology uses PFTs (Plant Functional Types) constant through the simulation period, assigning a type to each cell that will determine parameters like the root depth, the minimal stomatal conductance (that will be increased by atmospheric factors) and the LAI (Leaf Area Index), that will affect the calculation of canopy resistance, transpiration and evaporation from the canopy surface. The transpiration is taken from the moistest level in the root zone.*"

The persistence induced by the GW component must somehow be related to its long residence time (as written p12, L27):

(f) Is there a way to quantify it, at least at first order? How does persistence link with the transmissivity of the GW system (it would be useful to give information on it, how is Ks estimated, based on soil texture? which effective thickness?) and the GW-river flux (Qr), for which some quantitative parameter values would also be useful.

Authors: As in the response to (a) and (b), after this and other reviewer's comments, more information has been added in Section 2.1 about the transmissivity and Qr flux and how they are calculated. We believe this has made the paper more consistent.

River flow scheme:

(g) It is said that river width is taken form HydroSHEDS, but this variable does not belong to the standard dataset (https://hydrosheds.org/pages/availability). Please be more specific.

Authors: A complete description of how the river parameters have been calculated (including the new Fig. 2) has been added to Section 2.2 in the revised document.

The simulations are forced by an atmospheric reanalysis, ERA-Interim, without any bias correction except for precipitation:

(h) The reported horizontal resolution is about 0.7° x 1°, but the authors should check L30-31 p5, since I don't see why the resolution would be fixed in latitude and varying in longitude, it's usually the opposite which is done if seeking for constant grid-cell areas, but on the other hand, a factor of two over IB seem excessive.

Authors: Thanks for pointed this out. The resolution of the driving ERA-Interim data is now reported in Section 2.3, from Berrisford et al. (2011), as: "*ERA-Interim is presented in a reduced Gaussian grid with approximately uniform 79 km spacing for surface grid cells*". And the actual LEAFHYDRO model resolution, to which driving data are interpolated, is reported in Section 2.5 as: "*The simulation domain is a Lambert-Conformal grid centered in the Iberian Peninsula with a spatial resolution of 2.5 km*."

(i) Precipitation is bias-corrected and downscaled to the 0.2° resolution. At 40°N (inside IB), the area of a 0.2°x0.2° grid-cell is a bit less than 20^2 km², thus includes 64 LEAFHYDRO grid-cells (2.5^2 km², cf. resolution introduced L6 p7, when presenting the simulations). This resolution mismatch should be discussed, as it can have an impact on validation performances.

Authors: Yes, of course. In fact, as mentioned earlier, we believe that this is the main reason for the underestimation of winter streamflow in our simulations. We discuss it in Section 3.2: "*There is a clear underestimation of the winter river flow by the model. Two factors contribute to this bias. First, a lack of precipitation in the forcing data, since the IB02 analysis dataset original resolution (0.2º) is coarser than our model simulations and the station density (7 km in Spain and 11.7 km in Portugal) is not sufficient to capture precipitation peaks due to orographic enhancement over the mountains, which is very pronounced in the northern cordilleras.*"

(j) Better meteorological forcing data sets probably exist in Spain, as the SAFRAN dataset of Quintana-Segui et al. 2017, containing all the variables required to force a LSM at the 5km resolution and 1-hourly time step, for 1979-2014. Else, WFDEI (Weedon et al., 2014) is a ready-to-use forcing data set, with bias-correction and downscaling to the 0.5° resolution, based of ERA-Interim. The submitted paper should include a justification for choosing ERA- Interim compared to other products, especially given that Gonzalo Miguez-Macho, co-author of the submitted paper, is also co-author of Quintana-Segui et al. 2017.

Authors: Yes, we are aware of such datasets. The decision of using ERA-Interim an IB02 was adopted at the time of conceptualization and set up for this study, when the other (newer) datasets were not available. Both mentioned datasets, however, and also the newly developed MSWEP dataset (Beck et al., 2017), have been more recently used with LEAFHYDRO simulations in a study focused on droughts: *The Utility of Land-Surface Model Simulations to Provide Drought Information in a Water Management Context Using Global and Local Forcing Datasets* (Quintana-Segui et al., 2019)

(k) I couldn't find the time step of LEAFHYDRO, and it is required for a modelling paper. If the model time step is shorter than the one of the forcing dataset, the downscaling should be mentioned.

Authors: The time resolution for resolving heat and water fluxes in the soil and at the land surface was 60 s. The time step for groundwater-streams exchange, groundwater mass balance and water table adjustment in the WT run is 900 s. We have included this information in Section 2.5.

The meteorological driving data are linearly temporally interpolated to the model time steps.

Initial WTD: section 2.4 is not crystal clear for me, and some rewriting is advisable. In particular, the order of what is done is hard to follow, and the reasons to do what is done are not justified:

(l) There seems to be three successive initial WTD estimates at three different resolutions (1°; 9 arc-sec; the 2.5-km resolution of the simulations) but I don't understand at all what relates to the last two resolutions in the explanations of L12-16 p6. Can you please clarify?

(m) Why using recharge from a model without GW (Mosaic LSM) at 1°? Why not relying instead on the FD version of LEAFHYDRO model? At L10 p 6, is Qsr surface runoff?

(n) If topography is very important for the WTD patterns (L12 p 6), why using a higher resolution for the initial WTD and not for the transient simulation? Is it a problem of a computing power?

(o) The differences between the initial WTD (EWTD from Fig2) with the mean WTD over the 10 years (Fig 7b) should be discussed.

Authors: Yes, we agree that Section 2.4, in the original version, would gain with some clarifications and we have rewritten it. Hopefully the reviewer and the readers will find it now clearer.

In response to particular comments:

(l) the 1 degree resolution corresponds only to the initial recharge dataset from the Mosaic LSM that is used to feed the 2D groundwater model (Fan et al., 2007). This 2D steady-state model runs at 9 arc-sec and produces initial conditions for wtd at this resolution, resolving hillslope gradients.

(m) We needed an initial guess for climatological recharge to feed the 2D groundwater model. We had tried other datasets and results using Mosaic recharge, despite being very coarse, gave the best skill scores in validation with point observations. Running LEAFHYDRO with FD would mean an extra step that we deemed unnecessary.

Yes, Qsr is surface runoff.

(n) Of course the 2.5km resolution for the LEAFHYDRO simulations is a compromise needed for lack of computational resources. This is explained in Section 2.5 (now renamed "Simulations set-up")

(o) The differences come mostly from the colorscale and resolution difference in the plots.

<u>The way to obtain the power spectra of Fig. 14</u> is not at all explained but a few words wouldn't hurt.

(p) Shouldn't the compared curves have the same integral if they are calculated from time series of the same length?

Authors: We used the intrinsic function "spectrum" from the MathWorks software, which computes the power spectrum of a given time series, and it is simply based on performing the Fourier transform. We have added a comment about this in the manuscript.

3. Quantification of results

(a) The difference maps are interesting since they reveal clear sensitivity patterns related to WTD. Yet, an important part of the results is about water budgets, and means of the differences over IB would be interesting. This can be achieved either on the maps, or in a summary Table.

Authors: Yes, we agree. This was the reason to add the colour bars below each plot in figures 9 and 10 (now 10 and 11), representing the mean anomaly value for the peninsula or the zoomed area.

(b) An important question is about the significance of the reported changes in front of variability (seasonal and inter-annual), which can be assessed using inference tests. With simulations of only 10 years, non-parametric tests are probably advisable, and another solution would be to extend the simulation period, with additional advantages for persistence and long-term memory analysis.

Authors: Unfortunately, the simulation period could not be extended due to computational limitations.

(c) The validation of the models should involve more quantitative criteria. In particular, Fig 5 shows that WT strongly underestimates observed river flow (written p9 L7). Since ET is higher in WT than FD, one would expect that the river flow bias is smaller with FD (less ET with the same precip means more runoff and river flow): is it what is found? By how much? It doesn't seem true for the Ebro based on Fig. 15, which is weird.

Authors: The presence of the groundwater reservoir in the WT simulation buffers out variations in climate. Even with the same ET in both WT and FD run, in dry periods there can be more baseflow input in the WT simulation, coming from groundwater. The opposite can also be true. In a wet year, there can be less baseflow in the WT run if the groundwater reservoir levels are low from a previous drought. The mean flow for a given period should be about the same if there is no trend in groundwater levels. But this is not guaranteed either, because there can be some extra store or depletion of water in the layer between the water table and the bottom of the soil column at 4m, which is not existent in the FD run.

In Fig. 15, there is a declining trend in the water table, albeit small. In shallow water table areas, the lowering water table might be partially sustaining ET; however, where the water table is deeper, the lowering groundwater store is sustaining streamflow. This trend explains why there is more water in the annual mean total streamflow in the Ebro in the WT run, and is common to the other Mediterranean basins, more affected by the drought. We will now discuss this issue, related to the relatively short period of simulation, in the revised manuscript. The point that we wanted to make with the figure is still true, though, as it is apparent from the comparison with FD run that groundwater sustains streamflow not only through summers, but also longer dry periods.

(d) Eventually, what is the best simulation if we try to combine several performance criteria (correlation and bias, and also RMSE and Nash efficiency, or KGE which directly combines these scores, cf. Gupta et al., 2009)?

Authors: In terms of simulating river flow, we obtain better metrics (Nash-Sutcliffe efficiency) with FD, but in terms of simulating surface flow exchanges, we have shown how a more realistic soil moisture in the presence of the groundwater influence will result in different surface-atmosphere coupling effects, which is ultimately the issue we are focusing on. As mentioned earlier, streamflow results are more complex to analyze, and a better score with bad forcing does not mean that the simulation is more physically realistic.

(e) P8 L27-28 claims that "the model's performance is reasonably good at shallow water table depth points, but significantly worse where the water table is deeper": I don't think it is supported by any figure or result.

Authors: With this statement we refer to the improvement in the wtd validation reported when we consider only shallow water table points as compared with all points, deep and shallow are considered. We have rewritten the paragraph where we provide this wtd validation results and the point should be clearer in the revised manuscript.

(f) P10 L7-8: can you prove/justify/quantify how "small" is the long-term upward flux in flat areas?

Authors: The flat areas we refer to present values between 0 and 150 mm/yr, or between 0 and 0.5 mm/day in the case of the seasonal plots, in both cases, closer to 0, than to the upper value of the range. Our point was to separate them from the river valley areas we mention in the following sentences, where faster lateral drainage due to the steepness of the terrain result in much higher upward flux values.

(g) P12, L3: "precisely where the correlation between soil moisture and precipitation are reduced": this is not obvious from Fig. 9, and additional diagnostics would be interesting to prove this conclusion.

Authors: This sentence refers to the "missing plot" of differences between right and center maps in the figure that the reviewer mentioned in the concern (1g). We think that once the error in the caption has been corrected the intended exercise of focusing on differences between both maps is easier for the reader.

4. **Structure and writing**

(a) In absence of land-atmosphere coupling, the title is not well supported for the "land-atmosphere interactions", and should be modified.

Authors: We did debate about this wording during the writing process. We have now changed the word "interactions" for "fluxes" in the title in the revised manuscript.

(b) The introduction is long and messy, and would benefit from serious reshaping. The discussion on the need for realistic water table simulations (L34 p2 to L5 p3) is not well articulated with the rest, and is actually contestable. Besides, it raises questions since the WTD and river flow simulated by LEAFHYDRO (section 3.1) are not particularly realistic, although not very bad either. The paragraph at L13-20 p3 is very general and seems odd when the introduction starts to present the specificities of the presented work (starting at "Our work, p3, L5). The last part of the introduction (p2 L21 to P4 L8) reviews Spanish hydrology, and finishes on irrigation. Eventually, the specific research questions of the paper are not clearly stated by the end of the introduction.

Authors: We have reshaped the introduction as we agree with this and other reviewers in that the paper would benefit from it. Now the structure is clearer and simpler. It still starts from general statements on groundwater interactions, soil moisture memory and observational evidences of groundwater-soil coupling. Then we review other modelling efforts. Then we introduce particularities about the Iberian Peninsula and literature on it. Finally, we introduce the research questions (that we did introduce in the first paragraph in the original submission) and the particularity of LEAFHYDRO calculating lateral drainage that makes it a candidate to tackle challenges presented earlier.

(c) The paper frequently refers to a "bimodal" memory of soil moisture induced by GW persistence, but this term "bimodal" is not very clear: why isn't it just normal memory? I urge the authors to define what they really mean.

Authors: The author is right in that the term memory should suffice, but we introduced the term "bimodal" to insist on the capability of the fluxes to go two-ways and the soil moisture to not only remember dry conditions but also wet conditions providing water to the system during dry years following wet ones.

(d) Consider gathering the validation of river flow and sensitivity of river flow to WT vs FD.

Authors: Responded above.

(e) Consider presenting Fig. 10 before Fig. 9, which makes a nice introduction to Fig. 10, and justifies why time correlation is analyzed on time series of annual means, while many papers in the literature consider time lags of months. The paper may insist on memory at pluri- annual timescale, which is quite novel in the literature surface-GW interactions.

Authors: Yes, this is a good point about changing the order. We originally considered that we should send the message of the water table affecting the annual correlations and then presenting the annual maps in Figs. 10 and 11 (now 11 and 12) for the reader to better appreciate these effects. Of course we could change them if the reviewer insists. Thanks for the tip, we have included "at pluri-annual timescale" in the title of Section 4.2.

(f) The conclusion is mostly a summary of what was just presented in sections 3 and 4. The summary part should be strongly shortened, to better highlight the main findings instead of again comparing % changes. Another advantage would be to leave space for a real discussion of the results, which is cruelly lacking.

(g) In particular, the results and the conclusion they support are likely dependent on the model, its assumptions, and the forcing datasets (meteorology, soils, vegetation). This must be said and leads to compare the conclusions of the paper to the literature.

(h) For instance, the underestimation of riverflow (Figs 5 and 15) means that ET is too strong in the model(s) or precipitation too weak: can't it create a bias in the sensitivity of ET to WTD? This should be discussed, in relationship with the quality of meteorological and vegetation forcing datasets, or the fact that irrigation is not taken into account (cf. p9 L31-32).

(i) The perspectives could also be developed... Are "coupled land hydrology-climate models" (p16, L29) the only to move forward?

Authors: We have divided Section 5 into "Discussion", where we have discussed the results maintaining our structure of groundwater influences on three parts of the land surface hydrology system and have added comparisons with external work cited in the introduction, and a shorter "Summary and conclusions" section at the end. Please see the revised version of the manuscript.

The attribution of the improvements to our WT run, and therefore suggesting that the ET enhancement is right during the dry season is based mostly in our wtd validation. Therefore a validation of ET fluxes with suitable data is, of course, another perspective for future work. We just were not able to do it for this paper for lack of data and did not think it was necessary to mention in the text.

We believe, as has been pointed out in the text, that the driving precipitation is biased low.

References (for those not cited in the discussion paper):

Gupta, Kling, Yimaz, Martinez: Decomposition of the Mean Squared Error and NSE Performance Criteria: Implications for Improving Hydrological Modelling; Journal of Hydrology, 377(1), 80-91, doi:10.1016/j.jhydrol.2009.08.003, 2009.

Quintana-Seguí, P., Turco, M., Herrera, S., & Miguez-Macho, G.: Validation of a new SAFRAN-based gridded precipitation product for Spain and comparisons to Spain02 and ERA-Interim, Hydrology and Earth System Sciences, 21(4), 2187–2201. doi: 10.5194/HESS-21-2187-2017, 2017.

Weedon, G.P., Balsamo, G., Bellouin, N., Gomes, S., Best, M.J. and Viterbo, P., 2014. The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data. *Water Resources Research*, 50, doi:10.1002/2014WR015638.
Westerhoff, R., White, P., and Miguez-Macho, G.: Application of an improved global-scale groundwater model for water table estimation across New Zealand, Hydrol. Earth Syst. Sci., 22, 6449-6472, https://doi.org/10.5194/hess-22-6449-2018, 2018.