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.

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.

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.
   (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).
   (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).
   (d) Fig. 6: R seems negative if downward, which seems odd for the flux which recharges the GW.
   (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?
   (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?
   (g) Fig. 9 is very noisy: could you add the difference between center and left panels?
   (h) Fig. 11: please add the lon/lat of the mapped area, either on the maps, or on the caption.
   (i) Consider merging top panel of Fig 5 and Fig12; same for Fig 5 and Fig 15. 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).

(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.

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.

The recharge calculation, 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.

(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)?

(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.

(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?

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.

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.

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.

(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 km^2, thus includes 64 LEAFHYDRO grid-cells (2.5^2 km^2, cf. resolution introduced L6 p7, when presenting the simulations). This resolution mismatch should be discussed, as it can have an impact on validation performances.

(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.

(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.

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.

The way to obtain the power spectra of Fig. 14 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?

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.

(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.

(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.

(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)?

(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.

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

(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.

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.

(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.

(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.

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

(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.

(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?

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.