Review “The suitability of remotely sensed soil moisture for improving operational flood forecasting” by Wanders et al

The manuscript describes a case study for assimilating discharge and satellite soil moisture observations into a Lisflood model for the Upper Danube river and evaluates the performance on the forecast quality (one year).

Main Comments

The manuscript reads like a feasibility study. The manuscript is in parts inaccurate and unclear and overstates and generalizes the conclusions (as maybe typical for a feasibility studies applied to one study area) too much. This is also due to the fact that the term EFAS is often used, while the use of the term Lisflood model for the Upper Danube is more appropriate. The manuscript raises more questions than it answers.

Abstract
page 13784
Line 2-4: Replace EFAS with Lisflood model for the Upper Danube (throughout the abstract)
Line 4-5: Remove line about EFAS not sure why this needs to be in the abstract
Line 14: Replace show by suggest
Line 14-25: I don think this remarks are valid for Q1sat so you can not generalize

Page 13785
Line 1 Again I think show is to strong (there are too many thing unclear/not understood/ explained etc and validation is very limited)

Introduction
page 13785

- Line 11: Maybe also good to mention several other forecasting systems used in Europe (England, Scotland, France, Switserland, Austria, The Netherlands, etc.)
- Line 12/14: I find the statement “national forecasting systems are often not sufficient and transboundary forecasting systems are required” strange. What is a transboundary forecasting system? What do you mean by often (which rivers)? As an example, several countries (i.e. The Netherlands, Germany) run a flood/flow forecasting system in which transboundary rivers are modeled and used to generate forecasts for their own national domain.
- Line 13/15: I think the EFAS system was not developed to full fill the need described here, but was developed in support of crisis management at the European level.
- Line 24: For Example. What do the authors mean here? This is a not an example related to the lines above on state updating. Please remove this sentence as it is not appropriate/relevant here

- Line 26: However, it is difficult to obtain these measurements in real-time in a way they can be used in EFAS. Can the authors specify in more detail what the problem is (in a way they can be used in EFAS?)? Is the data not available in real-time? Or is there another issue? Please clarify

page 13786

- Line 11: The revisit time is 1-3 days how does this relate to the availability or usefulness of the discharge observations (in a way they can be used?)?

Page 13788

- Line25: One of the most difficult steps in a data assimilation setup is to determine the input and model uncertainty (PQR problem, see also Liu et al., HESS 2012 also for other relevant references). Here a setup is chosen which make use of 300 parameter sets, without any consideration what has been used by others even though this is the most critical step in the whole process. Because I don't know the study by Wanders et al. 2013 WRR I really cannot judge if this is correct, what the consequences are, etc I assume the bucket sizes vary between the different parameters set? How is this handled? Do you make use of maximum and minimum bounds in the data assimilation scheme? It is clear that this choice/assumption requires much more justification.
- Is this also the way the operational system is being envisaged to run?
- Is it correct that per setup Q0, Q1, Q7 300 different parameter sets are being used at least this is the way I read it and if not I don't understand the setup? What is the difference between those sets in terms of states/spreads/correlation in space/time etc.
- Line 5/6 "while the current EFAS uses fixed initial conditions for the hydrological forecast" I assume this is not correct or does EFAS use fixed initial conditions for each hydrologic forecast?

Page 13790

- Line 17-26: Here the assumption is made that the satellite soil moisture measurements (with 5cm or 2cm depth support) can be used for comparison with the Lisflood soil moisture bucket (theta_WP / theta_FC of the topsoil? How deep is the top soil?) through some scaling. What is the foundation/rational for doing so?

Page 13791

- Line5-10: Similar experimental setup using multiple interior discharge observation stations was used by Rakovec et al. 2012 and Lee et al 2012 (see http://www.hydrol-earth-syst-sci.net/16/2233/2012/hess-16-2233-2012.pdf)

Page 13793

- The experimental setup is not very clear given the objectives of the manuscript.

- The effect of data assimilation is very much depending on the perturbation of the model (already mentioned above)...
- Is the operational setup also based on running the model with 300 parameter sets? Or is the setup here ad-hoc or opportunity based? If the latter it is probably better to mention this upfront instead of presenting it as they way the operational setup will work;
- Why was the choice made to run the forecasts over the calibration period (not knowing what the calibration entails) a more independent testing of the setup seems more appropriate.

- Line 5 The error on the discharge measurement is set according to expert knowledge. How did others treat this uncertainty their DA setup. Maybe better to refer what has been used and state why and if the authors deviated from this.

Page 13794

- Line 1-10: By using only two (random) parameter sets the remaining initial condition uncertainty is basically removed in the forecast, in other words after the analysis we fall back to a more or less a deterministic model in forecast mode driven by EPS? Why 2 random parameter sets? And not 1, 4 or 6 etc?

Page 13795

- Q90 and Q80 are chosen for evaluation? How many Q90 and Q80 events (for one measurement location) are included in this one year period?

Page 13797

- I am interested understanding what happens to the model states also in spatial sense. Can you provide insight into where what happens to which model state in spatial sense?
- What is the reason why the flood peaks are overestimated for the Q1sat case? What happens there (also in spatial sense to the updated model states)?
- Why are all Skill Scores for the Q1sat case lower than for the Q1 case this seems very unlikely when looking at Figure 3 and Figure 4 especially for Q80/Q90?
- Does averaging the results over the (scaled) measurements give a false sense of accuracy as the events are the same? How statistically significant are the results per location/leadtime/threshold? Is the hindcast period not to short and does it contain enough model-observation pairs to justify any statement for higher thresholds?
- When zooming in on Figure 4 I see that for several Q1sat is shifted away from the discharge observation at T0 (at the analysis) what happens there? What is the reason for the large spread with Q1sat in the forecast? Even when the peak discharge arrives at T0 (last panel), state updating is not able to draw the model towards the observation. What is the reason for this? What happens here?
-

Conclusions page 13799

- Line 24/25 remove EFAS and use Lisflood model it suggests that results are valid for the modelled area (Upper Danube) .

Conclusions page 13800

- Line 8 “We show that the assimilation of remotely sensed soil moisture improves the flood forecasting especially” Is this true I would state “Our results suggest that the assimilation of remotely sensed soil moisture improves the flood forecasting only”. “we show” is, I think, overstated and I would use “suggest” (throughout manuscript and abstract)
- Line 14/15 This is not true Q0sat gives worse results (see results/figures etc)
- Line 19-27 Complete unclear. “This will ensure that the parameterization of the model is optimal for the correct simulation of the hydrological variables used in the assimilation framework” Was this shown? Is it an assumption by the authors? It is certainly not a conclusion in my eyes.

Page 13801

- Line 10-18 why coming with these results in the conclusions please move to the results section
- Line 19-27 This only valid for this Lisflood model for the Upper Danube, this model calibration, and the input data used and this setup. If the authors would use a model that was not biased, used different (better?) input data, the results might be completely different
- Line 20 replace EFAS by Lisflood model for the Upper Danube
- Line 21-23 “The addition of remotely sensed soil moisture will reduce the number of false positive flood alerts and thereby increase the reliability of the flood awareness system.” Again only when many discharge observations are being used and these are not available in real-time in a way they can be used in EFAS (see introduction)

Overall

The authors do state in the introduction that “it is difficult to obtain these measurements (discharge) in real-time in a way they can be used in EFAS” (maybe the same holds for the satellite data). Given the fact that the discharge observations are not usable, the authors remain very positive about the use of satellite data even when the results are not positive (and do not explain why) when only satellite data are used. I would expect a more balanced conclusion.