Hydrol. Earth Syst. Sci. Discuss., 9, C765–C769, 2012
www.hydrol-earth-syst-sci-discuss.net/9/C765/2012/

Hydrology and Earth System Sciences Discussions

Interactive comment on "Filling the white space on maps of European runoff trends: estimates from a multi-model ensemble" *by* K. Stahl et al.

Anonymous Referee #3

Received and published: 10 April 2012

General comments
In this paper, the authors evaluate the ability of an ensemble of large-scale hydrological models to reproduce observed trends in European hydrological regimes. Overall, this analysis is of great interest for the hydrological community. In particular, the spatial extent, the number of catchments and the number of models used in this assessment are quite remarkable and enable a meaningful evaluation. Moreover, the paper is well-written, the presentation of results is clear and the discussion is interesting. Consequently, I am definitely supportive of the publication of this paper in HESS. I have a couple of comments that should be addressed before publication, but since they mostly involve additional discussion, the resulting revision should be relatively minor.
My main comments are the following:

1. I think the title of the paper slightly misrepresents its actual content. In my opinion, the primary topic is the evaluation of the ability of models to reproduce observed trends (and the paper definitely fulfills this objective). "Filling the white space" is somehow another issue, and I don't think it is thoroughly tackled in this paper. Basically, "filling the white space" is achieved by replacing a map of sparsely observed trends by a map filled with modeled trends. The issue with this approach is that it also fills "non-white pixels", i.e. pixels where an observed trend is actually available! Is it reasonable to replace an observation by a simulation from one or possibly several models? Of course the answer to this question strongly depends on the ability of models to reproduce observed trends, so in this respect the analysis carried out by the authors is a necessary preliminary step. But I miss a discussion on how such a filling should be implemented. In my opinion, rather than simply replacing observed trends by modeled trends, it would be more reasonable to derive a composite map, using BOTH observed and modeled trends. This issue sounds kind of similar to the issue of merging radar and raingauge information (although the contexts are clearly distinct): a radar image is a spatially complete but inaccurate information on rainfall (as is a map of modeled trends in this study), while raingauges provide more accurate but sparse information (as do observed trends in this study). Merging both sources of information involves studying the spatial properties of radar-raingauge discrepancies (similarly to what is done in this paper in terms of discrepancies between modeled and observed trends), characterizing sub-grid variability and then using some geostatistical method to interpolate in-between raingauges (what I would qualify as "filling the white space"), while preserving the information at gauged pixels (up to sub-grid variability). See e.g. Severino, E., and T. Alpuim (2005), Spatiotemporal models in the estimation of area precipitation, Environmetrics, or Kirstetter, Delrieu, Boudevillain, Obled, 2010. Toward an error model for radar quantitative precipitation estimation in the Cévennes–Vivarais region, France, Journal of Hydrology, among many others, for illustrations. Note that I'm not suggesting that the same methods should be blindly applied, nor that such application should be performed within the present paper. But I believe a discussion

on the interest of merging information brought by the data and the models, rather than simply replacing the former by the latter, would add value to the paper. Moreover, the title could be more focused on the primary topic of the paper – namely, the validation of modeled trends.
2. The WFD forcing data used in this study are not direct measurements, but rather an interpolated and bias-corrected version of the ERA-40 reanalysis. In some sense, this is also a model. Consequently, I would find it interesting to evaluate whether the trends detected in WFD are consistent with the trends observed at weather stations and raingauges. Has such an evaluation been performed as part of the derivation of the WFD dataset? If so, it would be interesting to summarize the findings in this paper. If not, a discussion on the interest of such an evaluation would be valuable. Indeed, part of the discrepancy between observed and modeled hydrological trends could result from discrepancies between WFD- and station-based-trends in forcing variables.
3. In my opinion, an important application of the model simulations derived by the authors would be to start rigorously tackling the difficult problem of trend attribution: are observed trends consistent with the expected response of the catchments to the evolution of climate drivers? (the latter evolution being possibly low-frequency variability, climate change or both). The authors are discussing this issue in the discussion section, based on general considerations on the evolution of climate forcings. But such general considerations are necessarily limited by the complexity of the relationship between forcings and runoff. I think the availability of both observed and modeled trends may be an opportunity to use rigorous approaches to trend attribution. There is currently an interesting paper on this topic in discussion in HESS: "More efforts and scientific rigour are needed to attribute trends in flood time series", B. Merz, S. Vorogushyn, S. Uhlemann, J. Delgado, and Y. Hundecha, Hydrol. Earth Syst. Sci. Discuss, 2012. I think the authors should have a close look at this paper, and discuss how their model simulations could be used to implement the attribution framework proposed in this paper. See also Hundecha, Y. and B. Merz (2012), Exploring the relationship between changes in climate and floods using a model-based

analysis, Water Resour. Res., doi:10.1029/2011WR010527, in press, and Renard, B., et al. (2008), Regional methods for trend detection: Assessing field significance and regional consistency, Water Resour. Res., 44, W08419, doi:10.1029/2007WR006268.

Specific comments
p. 2008 line 12: accurate rather than reliable?
p. 2010 lines 10-16: (i) An histogram of catchment sizes, or at least more detailed statistics on catchment sizes, would be valuable; (ii) Is it frequent to have more than one catchment per grid cell? If so, it could be an opportunity to study the sub-grid variability of observed trends. It might be worth a short discussion; (iii) How many catchments are larger than the grid cells? What are the pros and cons of keeping/discarding such catchments from the analysis? A few words of discussion would be valuable.
P. 2010 lines 22-25: I find this statement quite strong, unless the bias discussed by the authors is constant in time and does not depend on the runoff-generating process (eg recession, flood, drought, etc). But the discrepancies between modeled and observed runoffs are most likely far more complex than such a constant bias, so that computing relative trends is not sufficient to ensure that the analysis is not affected by the discrepancies between observed and modeled runoffs. I think this sentence should be qualified or reworded.
p. 2011 line 14: dependence in space is not an issue to determine the significance of at-site tests – it is for regional testing procedures.
P. 2012 lines 9-13: some authors proposed quantitative metrics to compare spatial patterns- it might be of interest to the authors. See e.g. Davis, C. A., B. G. Brown, and R. G. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. Mon. Wea. Rev., 134, 1772–1784.
P. 2013 lines 17-22: it's a pity not showing results for monthly runoffs – maybe a synthetic figure using eg boxplots would be feasible?

P. 2013 line 22: “significant similar distributions” sounds awkward within the terminology of statistical tests – the distributions can only be “significantly different”, otherwise they are “not significantly different” rather than “significantly similar”.
P. 2015 lines 13-14: this sounds like an interesting result, since it suggests that model agreement is an indication of the accuracy of modeled trends (which wasn’t a priori obvious... models could agree but be all wrong!). I’m wondering whether this could not be diagnosed in further depth, e.g. by splitting the spatial domain in several groups according to the level of agreement between models, and quantifying the discrepancies between observed and modeled trends for each group.
P. 2017 lines 6-8: This statement might be a bit too strong – it only holds for forcing variables the models are sensitive to for annual runoff and high flow. But it doesn’t prove the reliability of all forcing variables, at any time scale.

Technical corrections
P. 2011 line 7: mm.yr-1
Figure 1: colors would improve readability, at least to better distinguish observations and ensemble mean.
Figures 4 and 5: the crosses and squares are really difficult to distinguish – maybe use a different set of colors or just increase figure size?

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 9, 2005, 2012.