Detailed response to the comments of Massimiliano Zappa

Our responses are shown below in blue; Mr. Zappa's comments are shown as normally black text.

Comment: Page 1, line 21: Spell maybe them [the performance indicators] out in the abstract.

Reply: We will explicitly mention the performance indicators used in our study. The sentence is now: "The performance of both forecast methods is evaluated in relation to the climatological forecast (ensemble of historical streamflow) and the well-known Ensemble Streamflow Prediction approach (ESP, ensemble based on historical meteorology) using common performance indicators (correlation coefficient, mean absolute error (skill score), mean squared error (skill score), continuous ranked probability (skill) score) as well as an impact-based evaluation quantifying the potential economic gain."

Comment: Page 1, line 23 – 25: Nice list. Finding 1) could be better formulated. Such as:"..... Europe indicate, the accuracy/skill of the meteorological forcing used has larger effect than the quality of initial hydrological conditions for relevant ..."

Reply: We will adopt your suggestion and change the sentence to: "1) As former studies for other regions of Central Europe indicate, the accuracy / skill of the meteorological forcing used has larger effect than the quality of initial hydrological conditions for relevant stations along the German waterways."

Comment: Page 2, line 19-20: In some rivers is possible to have 2/3 small vessels instead of a big one causing increasing transport costs, but no reduction of the amount of goods transported?

Reply: That's correct and it is detectable that in case of low flow situations the total number of ships, e.g. on the River Rhine, increases noticeably. In order to clarify this aspect, we will add the following sentences: "During low flow situations additional (small) ships are needed to handle the same amount of cargo as during periods with mean water-level conditions. This cause increasing transport costs (see figure 1). But especially in the case of extreme or long lasting low flow periods the available small ship volume is limited compared to the transport demand and goods have to be shifted to other modes of transport, if this is technically possible free and if transport volume is available there."

Comment: Page 3, line 16: In this paper we do not focus on seasonal forecasting, but we present a kind of ESP. Our recent work on monthly predictions is presented here:
- Jörg-Hess S, Griessinger N and Zappa M. 2015. Probabilistic Forecasts of Snow Water Equivalent and Runoff in Mountainous Areas. J. Hydrometeor, 16, 2169–2186. doi: http://dx.doi.org/10.1175/JHM-D-14-0193.1
- Fundel F, Joerg-Hess S, Zappa M. 2013. Monthly hydrometeorological ensemble prediction of

streamflow droughts and corresponding drought indices. Hydrol. Earth Syst. Sci., 395-407, doi:10.5194/hess-17-395-2013.

Reply: We will add the two references suggested.

Comment: Page 3, line 5: Consider linking your paper with EFAS:
- Bartholmes, J. C., Thielen, J., Ramos, M. H., and Gentilini, S.: The european flood alert system EFAS – Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts, Hydrol. Earth Syst. Sci., 13, 141-153, doi:10.5194/hess-13-141-2009, 2009.

As stated in a short comment also the efforts of Demirel et al. should be cited: Demirel, M. C., Booij, M. J., and Hoekstra, A. Y.: The skill of seasonal ensemble low-flow forecasts in the Moselle River for three different hydrological models. Hydrol. EarthSyst. Sci., 19, 275–291,2015

Reply: We will add Demirel et al. 2015 as citation. Additionally we will create a link to EFAS based on the recently introduced "seasonal hydrological outlook" offering hydrological forecasting information up to two months ahead. With this outlook EFAS offers Europe-wide homogenous information on surplus or deficit of water resources. We will add Bartholmes et al. 2009 as reference and we will include EFAS in the discussion section with respect to the presentation of seasonal forecast results as well as the way forecast skill is assessed based on multiple verification criteria. We will add: "Since 2016 the European Flood Awareness System (EFAS, Bartholmes et al. 2009) offers a seasonal hydrological outlook with a lead-time of two months, too. This outlook offers Europe-wide homogenous information on surplus or deficit of water resources as weekly averages aggregated for major hydrological units. Still the generation of flood early warning information up to two weeks ahead for regional, national or European authorities remains EFAS focal point, but the seasonal products support this continental hydrological information service."

Comment: Page 5, line 17: How do you compute the mean low flow?

Reply: The flow statistics are based on the "Deutsches Gewässerkundlichen Jahrbuch"; the mean low flow is calculated as the arithmetic mean of the lowest daily flows of each year within the reference period. This explanation is added to the text. Furthermore we will include the reference periods used to calculate the mean / the mean low flows for each gauge in table 1.

Comment: Page 6, line 17: Consider to discuss bias-correction using this recent paper: How Suitable is Quantile Mapping For Postprocessing GCM Precipitation Forecasts? TONGTIEGANG ZHAO et al. J. of Climate, http://journals.ametsoc.org/doi/pdf/10.1175/JCLI-D-16-0652.1

Reply: Bias-correction is an important topic also in the context of seasonal forecasting in order to use the (coarse) output of GCM's at the (smaller) spatial scale of most hydrological models. But the influence of different bias correction methods is not in the scope of this paper. We already worked somewhat on this topic (and we will enhance our effort in this regard in the future), but so far we didn't detect significant improvements by more sophisticated methods. That's why we applied the most simple and hopefully most robust one in the context of this study. Undoubtedly it is a relevant aspect; therefore we will add some additional explanation and references on this topic in the revised paper:

- Page 6, line 17 ff: "...the data from E-OBS and ERA-Interim had to be downscaled and bias-corrected. A large number of different bias correction methods are available ranging from simple methods as linear scaling (see e.g. Lenderink et al. 2007), linear scaling with additional correction of the standard deviation (Leander and Buishand 2007), to more complex methods such as the distribution based correction method Quantile Mapping (se e.g. Piani et al 2010). We applied these different methods to bias correct E-OBS and ERA-Interim and compared the goodness-of-fit of the flow simulation using the downscaled and bias-corrected meteorological input. As the goodness-of-fit measures such as the NSE and correlation were very similar we decided to use the most simple bias correction method linear scaling. Linear scaling with reference to the HYRAS dataset was applied on a coarse grid (25 km x 25 km) for each month...."
- Page 7, line 2ff: "Before feeding the hydrological model, the output from S4 (daily total precipitation and air temperature), interpolated to a 50 kmx50 km grid (multiple of the 5 km x 5 km model grid), was bias-corrected with the meteorological observation dataset used for the baseline simulation. Again several bias correction and post-processing methods of different complexities for ensemble forecasts are available (see e.g. Crochemore et al. 2016, Zhao et al. 2017). After the experiences of the bias correction of EOBS and ERA-Interim we decided to stick to the most simple bias correction method linear scaling successful applied for bias correction of seasonal forecasts (Crochemore et al. 2016). In future applications different bias-correction and post-processing methods will be applied and analyzed."

Comment: Page 8, line 2: Are they subdivided or not in your specific application?

Reply: This relatively new option in LARSIM is already used for the Rhine basin, for Elbe and Danube the models have to be updated referring to this. We will clarify this aspect in the paper by adding this paragraph: "To avoid unnaturally high snow accumulation and to consider snow mass transport mechanism the new LARSIM snow mass transport option could be activated. Using this option snow only can accumulate up to a gradient-dependent threshold and exceeding snow is simply passed to the next model element downhill. These options (subdivision in elevation zones and snow mass transport) in LARSIM are currently used for Rhine basin. For Elbe and Danube they will be implemented in the future."

Comment: Page 8, line 6-8: One wish to learn more about this step. There is some experience in regionalization of parameters in such kind of watersheds (Viviroli et al., 2009, http://www.sciencedirect.com/science/article/pii/S0022169409005186)

Reply: We chose the following clustering and regionalization approach (added to the revised text):

1) Definition of statistical flow values for the 132 headwater catchments. As statistical values we chose
   - Mean flow / basin area
   - High flow with a 2-year recurrence frequency / mean flow

  - Monthly mean low flow / mean flow
  - Mean flow in winter / mean flow in summer
2) Identification of 9 clusters using the k-means clustering algorithm
3) Selection auf 7 geographical factors (e.g. height, slope, areal share of forest, areal share of field, areal share of unconsolidated rock, mean permeability of upper and lower soil) in order to characterize the clusters
4) rule-based mapping of all subbasins to the clusters

Comment: Page 8, line 9: I am not totally against manual calibration, but the authors should justify their choice here. Additionally, here I have some standard questions: Why do not chose one of the many ways to make automatic calibration? Can you comment on equifinality? How do you deal with snow towers in the Alps? Which parameters have been varied? How was the parameter range adopted?

Comment: Page 8, line 11-12: I need here some supplementary information on calibration and VALIDATION.

Reply: We will add the following paragraph (to page 8, line 9) in order to explain the calibration / validation strategy and to justify our choice for manual calibration:

"The manual calibration for the period 1998 to 2006 (validation 1976 – 2006) followed the guideline to calibrate LARSIM water balance models (Haag et al. 2016), which recommends the relevant parameters to be calibrated, the parameter range and a calibration procedure. The manual calibration strategy applied involved the following steps: 1) adjustment of the parameters of the snow modules with focus on the water balance and floods caused by snow melting, 2) adjustment of base flow storage relevant parameters to reproduce discharges at low flow conditions, 3) adjustment of the parameters relevant for interflow to reproduce mean flow conditions, 4) adjustment of relevant parameters to reproduce flood hydrographs 5) final validation and fine adjustment of all parameters. By using this well-proven process-based calibration strategy for LARSIM models conducted by experienced hydrologists instead of a non-process based automatic calibration procedure we expected to reduce the degree of parameter uncertainty in the clusters due to the problem of parameter equifinality (Beven 1996). But the fact that different sets of model parameters reproduce equally good output signals stays an issue in any hydrological model calibration and it is a significant aspect in particular when applying those models for predictions. Afterwards the parameter means and parameter spans have been derived for the clusters and transferred to the respective clusters in the whole model domain. As a next step a fine calibration of the model parameters within the parameter spans of the clusters have been conducted for selected parts / gauges of the Upper Danube, the Elbe and the River Rhine. The fine calibration for Elbe, Danube and Rhine was based on the same period an input dataset (HYRAS). Special attention was given to anthropogenic effects dominating the flow behavior in several catchments (dams, regulated lakes, water transfers) and the most relevant structures have been implementing explicitly. In order to retain the consistent spatial parameter distribution, the parameters have been optimized within the spread of the specific cluster of in the initial calibration. The standard deviation of the parameter values has been used as indicator to identify potential room for parameter optimization.

In the forecast framework different meteorological input (EOBS, ERA-Interim instead of the non-operational HYRAS-dataset) has to be used. Therefore we show in Figure 3 the output from the LARSIM-model set-up for monthly to seasonal forecasting for the hindcast period from 1981 to 2015."

- Haag, I., Johst, M., Sieber, A., Bremicker, M. (2016): Guideline for the Calibration of LARSIM Water Balance Models for operational Application in Flood Forecasting–Calibration Guide, 03.03.2016. http://hmdblog.rlp.de/luwg/larsim/

Comment: Page 8, line 15: This is nice, but not that challenging at monthly scale. Do you have data also at the daily scale?

Reply: Of course, you're right. We chose showing the model performance on monthly scale as we aggregate our forecast information on monthly averages, too. But also the daily model performance is good, we would claim (see table below). Along the Elbe river there are still some issues mainly with the routing. A comparison of the monthly and daily model performance is shown in the following table:

	NSE monthly	NSE daily	r monthly	r daily
Kaub	0.81	0.81	0.94	0.93
Hofkirchen	0.88	0.86	0.95	0.93
Neu Darchau	0.87	0.73	0.94	0.87

We will add to the paper the following sentence: "On the original daily time-step, the gauges show NSE > 0.7 and r > 0.85. Especially for the Elbe the model performance on daily basis is somewhat lower, mainly due to shortcomings in the river routing."

Comment: Page 8, line 17: You should spend some words on how hydropower management and lake regulations are accounted. Do you have also data for sub-monthly discharge (e.g. daily data)?

Reply: In the current set-up of LARSIM_ME most of the major barrages, regulated lakes and anthropogenic water transfers have been implemented. The known rules, as far as they have been provided by the different water management authorities, were implemented, too. Of course, this is a significant simplification and we know from the short- to medium-range forecasts, that the real regulation often differs from theoretical regulation rules. The completion of anthropogenic storages / transfers for the models of different basins is an important future task, not solely relevant for seasonal forecasting.

E.g. in the Alpine part of the LARSIM model of the Rhine basin 7 lakes are implemented (upper and lower part of lake Constance, Lake Brienz, Lake Thun, connected lake system of Lakes Neuchâtel/Biel/Morat, Lake Sempach, Lake Lucerne, Lake Zürich). Additionally 5 water transfers are been implemented so far. We don't have daily data available in order to validate in detail the regulation of the different dams / lakes.

In the paper we will relate to this aspect: "Another effect is anthropogenic influences, especially regulated dams, lakes and water transfers. As it is quite challenging to account for such impacts in a large scale hydrological model like LARSIM_ME, just the major barrages and water transfers together with their regulation rules have been implemented so far. This is not due limitations in the model functionality, but due to the difficulties in getting the required information. Therefore some of the problems the LARSIM models show to reproduce the flow behavior result from missing or incomplete reproduction of anthropogenic effects."

Comment:Page 8, line 18: See main comment. Here you can expand in the discussion and consider assimilation: Jörg-Hess S, Griessinger N and Zappa M. 2015. Probabilistic Forecasts of Snow Water Equivalent and Runoff in Mountainous Areas. J. Hydrometeor, 16, 2169–2186. doi: http://dx.doi.org/10.1175/JHM-D-14-0193.1 and rerferences therein

Reply: We will included this aspect in the new discussion section.

Comment: Page 9, line 9: References?

Reply: We will add Yossef 2013, Svensson 2016 and Tucci et al. 2003 as references proving that monthly and 3-monthly flows are quite common forecast parameter:

- Tucci, C. E. M., R. T. Clarke, W. Collischonn, P. L. da Silva Dias, and G. S. de Oliveira, Long-term flow forecasts based on climate and hydrologic modeling: Uruguay River basin, Water Resour. Res., 39(7), 1181, doi:10.1029/2003WR002074, 2003.

Comment: Page 9, line 10: References?

Reply: We will add two references regarding the use of NM7Q:

- Marke, T.: Development and Application of a Model Interface to couple Land Surface Models with Regional Climate Models for Climate Change Risk Assessment in the Upper Danube Watershed. Dissertation, LMU München: Fakultät für Geowissenschafte. 2008.
- Richter, B. D., Baumgartner, J. V., Braun, D. P., and Powell, J.: A spatial assessment of hydrologic alteration within a river network. Regul. Rivers: Res. Mgmt. 14, 329-340, 1998.

Comment: Page 10, line 2: Large previous work using this caveat. Please add references.

Reply: We will add as references Wood et al. 2016 as well as:

- Shukla, S. and Lettenmaier, D. P.: Seasonal hydrologic prediction in the United States: understanding the role of initial hydrologic conditions and seasonal climate forecast skill, Hydrol. Earth Syst. Sci., 15, 3529-3538, doi:10.5194/hess-15-3529-2011, 2011.
- van Dijk, A. I. J. M., J. L. Pena-Arancibia, E. F. Wood, J. Sheffield, and H. E. Beck: Global analysis of seasonal streamflow predictability using an ensemble prediction system and observations from 6192 small catchments worldwide, Water Resour. Res.,49, doi:10.1002/wrcr.20251, 2013.

Comment: Page 11, line 21 - 23: I think you changed colors on the map but not in the text, green should read violet and yellow should be orange, isn't?

Reply: You're right! Indeed we missed adapting the text. The correct sentence is: "Regions / grid cells showing a positive, stable correlation will be represented as red (90%) / orange (80 %) on a global map, while areas with a negative and stable correlation will be represented as blue (90 %) and violet (80 %)."

Comment: Page 12, line 11 – 12: To be fixed ... I am not sure whether a table is missing here. I think you refer

Reply: Sorry, stupid mistake. We will also correct the sentence slightly. It should be: "The last column of Figure 11 shows the forecast results without using measured discharges as predictors …"

Comment: Page 13, line 19: Eg. also: Fundel F, Zappa M. 2011. Hydrological Ensemble Forecasting in Mesoscale Catchments: Sensitivity to Initial Conditions and Value of Reforecasts. Water Resour. Res., 47, W09520, doi:10.1029/2010WR009996.

Reply: Thank you for this note. We will add (Fundel and Zappa 2011) as reference for a hydrological application of the relative economic value.

Comment: Page 13, line 21-24: Good choice. Very challenging is the setup of a user-tailored C/L ratio. In case of navigation this could be feasible, in my opinion. Do you have data to work on that?

Reply: Thank you for this interesting comment. We're currently working in the H2020-project IMPREX in close interaction with several stakeholders, and we hope to identify realistic C/L ratios, but so far, we don't have this information.

Comment: Page 15, line 10-12: We experienced that good initial conditions are important for low-flow periods. Demirel et al. (2013) discusses also the topic.

Reply: That coincides with our findings. We will add the two references suggested: Fundel et al. 2013 and Demirel et al. 2013.

Comment: Page 15, line 13: Have you thought about also including information on SNOW to predict low-flow?

Reply: So far, snow is just implicitly included in the forecasting process as it is modelled by the hydrological model. Indeed we're working on explicitly use data / information on snow (snow coverage, snow water equivalent) quantitatively by assimilating this information into the hydrological model. This is a relevant aspect for our short- to medium-range low flow forecasts, too. As data sources we're looking at measured and modelled data from DWD (SNOW4-products) as well as remotely sensed data from the EUMETSAT-funded H-SAF project (http://hsaf.meteoam.it/), see e.g.

- Alvarado-Montero, R., Schwanenberg, D., Krahe, P., Lisniak, D., Sensoy, A., Sorman, A. A., Akkol, B.: Moving horizon estimation for assimilating H-SAF remote sensing data into the HBV hydrological model. Advances in Water Resources 92 (May), 248-257, 2016.

We included this aspect in the new discussion section.

Comment: Page 15, line 26-27: Please more discussion on this. Some skill is hidden in the snow.

Reply: We will pick this topic up in the new discussion section.

Comment: Page 18, line 18: Discuss

Reply: We will add the following sentences in order to explain the background of this finding: "The S4-driven forecasts show the highest hit rate for all events, but especially for the highest threshold (50^{th} perc.), the false alarm rate is quite high (0.59), too. In this case the median don't seem to be the optimal representative for the ensemble and choosing another quantile might lead to better economic vales. Also the ESP-forecast suffer from a comparatively high false alarm rate. For the more rare low flow events, the false alarm rate of the S4-driven forecast significantly drops and is relatively close to the one from ESP- and MLR-forecast, while the hit rate stays best."

Comment: Page 18, line 22: Why changing the thresholds from Figure 9 to Figure 10?

Reply: We decided not to use the recurrence intervals (shown in figure 9 for different approaches at the River Rhine) when inter-comparing the statistical forecast at different rivers (figure 10) in order to avoid considerably differing numbers of events. The use of percentiles makes sure that the number of events to be forecasted is identical for each river.

Comment: Page 19, line 6: Maybe put also in the Figure that this numbers are for MLR

Reply: Yes, we will extend the legend in figure 10 in order to make clear that the numbers are based on the MLR approach

Comment: Page 20, line 7: After your submission this paper was also accepted: Laaha et al. 2016

Reply: Thank you very much for this hint; indeed an interesting paper well fitting in this topic. We will add this paper as reference.

Comment: Page 24, line 5: I expect in the conclusions the enumeration of the 4 findings presented in the abstract.

Reply: We will integrate our four main findings explicitly in the abstract.

Comment: Page 29, line 22: Switched down by one place in the alphabetic order

Reply: We will update / correct the alphabetic order of the references.