Detailed response to the comments of the anonymous referee #2.

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

Major comments

MoNM7Q is used to define the low flows each month. So the authors focus on 1 value/month. What about longer flow deficits or if several low flows occur during the same month?

Reply: The use of MoNM7Q as forecast parameter for longer lead-times is based on discussions with navigation-related forecast users. This parameter is known as a common and robust low flow indicator often applied on an annual basis in order to characterize the low flow situation of a year. In our case this parameter is used to characterize the monthly (low) flow conditions. As you mentioned, MoNM7Q doesn't provide intra-monthly information (e.g. timing of the lowest flows within the month etc.). But for decision making in the IWT sector on a monthly or seasonal time-scale this temporal resolution is sufficient, according to the users we're in contact with. This is comprehensible as the users don't have to estimate the load of a specific ship on a specific trip weeks or months ahead, but a typical question is, for example, how to compose an optimal fleet for the coming month(s). Therefore it is feasible to rely on more aggregated hydrological information like the "average" or the "worst" conditions expected, characterized e.g. by the mean flow (MQ) or the lowest arithmetic mean of 7 consecutive daily values (MN7Q).
Furthermore the flow dynamic of the free-flowing waterways Rhine, Elbe and Danube is relatively low. Therefore low MoNM7Q values in these rivers require a dry period of several preceding months (as mentioned in line 10/11 on page 9). Although it might happen that a low flow situations is temporarily interrupted by several days of moderate conditions (due to intensive rainfalls in one of the larger subbasins), two hydrologically independent low flow events won't arise within one month.

In order to clarify that forecast information on a monthly resolution is sufficient for decision making on longer lead-times, we added the following sentence to section 3.2: "For decision making in the IWT sector on monthly or seasonal time-scales a monthly / three monthly resolution is sufficient. Users don't estimate the load of a specific ship on a specific trip, as on short- to medium-ranges time-sales, but typical decisions (e.g. how to compose an optimal fleet for the coming month) require information on "average" or the "worst" conditions characterized e.g. by the monthly mean flow or the lowest arithmetic mean of 7 consecutive daily values within a month."

The stability correlation map is not clear. How the correlation is calculated. If this procedure aims at identifying predictors that should be calculated with a lag (monthly mean SST vs. streamflow month +1). This is not mentioned.

Reply: The stability correlation map is a tool used to identify stable and respectively suitable predictors for the statistical forecast approach. A predictor is defined by the variable (e.g. SLP), the region (e.g. North Atlantic Ocean) and the lag (e.g. -1 month). Based on the following three steps stability correlation maps are generated:

1) Each predictand (e.g. 3MoMQ for March-April-May at station Kaub / Rhine) is correlated with numerous potential predictors. Different lags (e.g. mean sea level pressure in February, mean sea level pressure in January) and regions of the same variable are regarded as independent predictor.

2) The correlation is computed in a moving window of 31 years within the period from 1948 to 2012 (window 1: 1948-1979, window 2: 1949-1980, …, window 34: 1981-2012).

3) The correlation is considered to be stable for those grid points / regions where predictor and predictand are significantly correlated at the 90% level and respectively 80 % level for more than 80% of the 31-year windows within the period 1948–2012. According to the level of significance, grid points / regions are colorized in red to yellow shades (stable positive correlation), shades of blue to green (stable negative correlation) or white (non-stable correlation) in order to create a stability correlation map for each potential predictor.

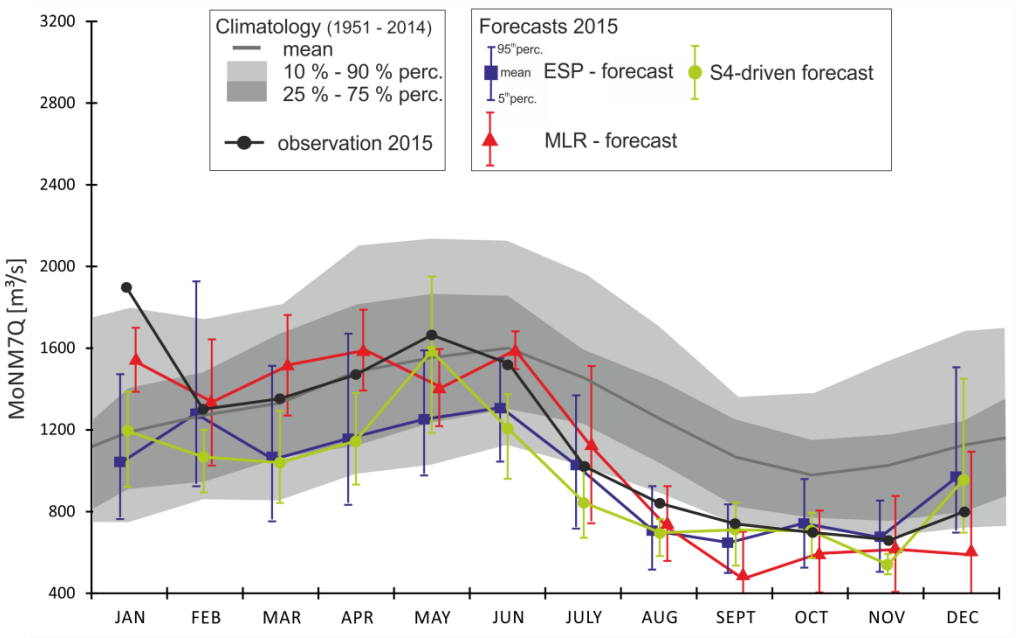
We will extend the description of the stability correlation maps in this way. Furthermore, we updated figure 4 in order to enhance comprehensibility and to show the real data used in this study. We will also add the following sentence to improve the description of figure 4:

"In the left part of figure 4, the correlation is plotted in the middle of each 31-year window. So, the first point represents the correlation between 3MoMQ and previous winter SLP from 1948 to 1979, while the last point represents the correlation from 1981 to 2012."

Seasonal S4 in an ensemble forecast whereas the statistical forecast is not (as I understand). It is difficult to compare. For instance, in the last section (case study). How the S4-driven forecast is defined in Fig. 12, 13, 14. It should be represented as a 'spaghetti plot' of boxplots. If the authors use a specific member, the median for instance, did they test different definitions before to be sure that was the most accurate?

Reply: You are totally right. The dynamical approaches (ESP and S4-based forecasts) are ensembles, while the statistical approach generates a deterministic forecast (including uncertainty bounds). In order to compare both types of forecast, we converted the ensemble forecast into a deterministic one by choosing the ensemble mean as representative of the ensemble. We think the use of the ensemble mean (as the central tendency of the ensemble forecast) is justifiable, because it is one, maybe the typical, standard index of an ensemble and the majority of forecast users, we are working with, choose this value to base their decisions on. So far, we (and the users) haven't tested alternative indices (e.g. different quantiles or specific members) of the ensemble to be used as deterministic forecast information. Of course, you're right, that by doing so, we lose some information originally provided by the ensemble.

Following your comment, we evaluated the forecast uncertainty represented by the ensemble spread additionally. As mentioned before, we have similar information from the statistical approach available. Therefore we defined as uncertainty range for the ensembles the range between the 5^{th} and 95^{th} percentile. Fig. 12, 13 and 14 have been updated accordingly. The forecast (as well as the climatology) are displayed like a boxplot indicating the mean and the 5^{th} and 95^{th} quantiles (see updated Fig. 12 below). In the text we picked this aspect up, too.



Some forecast assessments are not well calculated. For instance P13L6, the score of the ensemble is calculated by averaging the score of each individual member. The purpose of an ensemble forecast is not to generate good forecasts of each member.

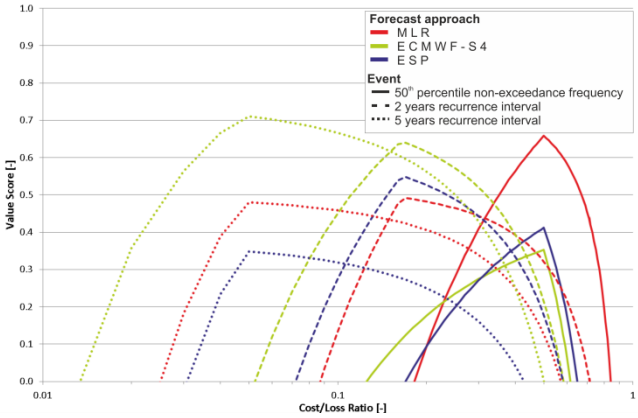
Reply: We decided to calculate the deterministic metrics MAE and MSE for each ensemble member first and calculate the mean afterwards. The idea behind this procedure is that besides assessing the accuracy we want to consider the spread / sharpness of the ensemble, too. The method we used rewards accuracy, but punishes spread. Of course, you're right with your statement and we agree that an ensemble forecast should be verified primarily probabilistically. But on the other hand each ensemble member represents an equally likely scenario of future conditions. As this way of generating the MAE / MSE is even stricter than treating the ensemble as one single trace (the ensemble mean), we would like to keep the corresponding results for the evaluation within this study.

In Results and Fig. 7, 8, and 11, it is not clear how the authors deal with ensemble forecasts. Which member is used and why? In the same part, most of the scores assess the capacity of the forecasts to represent the intra-seasonal variabilities not necessary the prediction of extreme events (MoNM7Q exits every month and is not necessarily climatologically low). Some assessment should be done focusing on dichotomous forecast of extreme events.

Reply: In general we used the ensemble mean as representing index of the ensembles throughout this study. As mentioned above, the ensemble mean is the information users from the transportation sector asked for in the vast majority of cases. But, as you mention this aspect, we should clarify this throughout the paper (e.g. by referring to chapter 3.4).

Most decisions in the context of IWT requiring monthly to seasonal forecasts – at least as far as we know so far – aren't typically based on thresholds. This might be one of the major differences with regard to flood forecast, where e.g. the level of protection is an obvious threshold. The logistic

managers have to optimize the transport capacity as good as possible with regard to the expected flow in the coming month(s), even if the flow conditions aren't extremely high or low. Against this background, dichotomous forecast evaluation is of limited relevance for the IWT sector and it doesn't play a major role in our evaluation. Nevertheless in order to validate the relative economic value of the different forecast approaches (see Fig. 9 and 10), we had to convert the (continuous) forecast into dichotomous ones. As thresholds / events we selected recurrence intervals of 2 and 5 years for the MoNM7Q at Kaub. The selection of even more extremes doesn't seem to be useful due to the limited validation period 1981 to 2014 (availability of System-4 reforecasts). In order to follow your remark and to be consistent with our existing evaluation, we plotted the probability of detection (or hit rate) and the probability of false detection (or false alarm rate) next to the economic value curves in Fig. 9 and 10 (see updated fig. 9 below). We think that's a meaningful addition as the maximum relative economic value score corresponds to the difference of the hit and the false alarm rate. Additional categorical statistics calculated internally (e.g. frequency bias, threat score etc.) don't provide additional information to benchmark the different forecast approaches.



	50th perc.		RI = 2 years		RI = 5 years	
	POD	POFD	POD	POFD	POD	POFD
ESP	0.89	0.48	0.64	0.09	0.38	0.02
ECMWF - S4 (dynamical aaproach)	0.94	0.59	0.75	0.11	0.75	0.04
MLR (statistical approach)	0.81	0.15	0.54	0.04	0.50	0.02

Value Score area	50th perc.	RI = 2 years	RI = 5 years
ESP	0.126	0.173	0.081
ECMWF - S4 (dynamical aaproach)	0.117	0.215	0.213
MLR (statistical approach)	0.293	0.205	0.166

A combined product, which uses both statistical and dynamical approaches as an ensemble, should be tested here to see what is benefit of using this advanced, and maybe the most accurate, method for the users.

Reply: What you suggest is exactly what we intend to investigate in more detail in the future. In section 5 we already showed the potential of using ESP-forecasts as additional predictor of the statistical approach. And, of course, other ways of combining the different approaches are possible and should be tested (keyword: hybrid forecast approaches). E.g. Olsson et al. (2016) evaluated various ways of combining forecasts based on a similar set-up. But before working on forecast combination we would like to test some post-processing techniques for the single approaches before spending time in combining the forecast of different approaches. That's why we can't serve with well-founded results of combined forecast results at the moment. But we'll add this topic to the new section 5 "discussion" (requested by the other referee).

Minor comments

P3l20: be careful order different to those in the abstract.

Reply: That's a good hint. We switched the order in this paragraph so that it corresponds to the one of the abstract.

P4l25-26: reference needed, please refer to Fig. 2 right panels here.

Reply: We added Funk et al. (2015) as reference, which describes the hydro-climatological conditions of the River Rhine waterway and which analyses critical situations for waterway transport. Additionally we included a link to figure 2 (right panel), as you suggested.

> Funk, D., Pouget, L., Dubus, L., Falloon, P., Meißner, D., Klein, B., Palin, E., Viel, C., Foster, K., Lootvoet, M., Bosi, L., Creswick, J., Davis, M., and Jimenez, I.: White paper on sector specific vulnerabilities. Deliverable 11.2, EUPORIAS - European Provision Of Regional Impacts Assessments on Seasonal and Decadal Timescales, Grant Agreement 308291, 2015.

Fig. 2: Quality of the images should be improved

Reply: We regenerated the single images of figure 2 and increased its resolution.

Table 1: Definition of the low flows not clear here.

Reply: We agree. This issue has been mentioned by the other referee M. Zappa, too. In order to clarify this aspect we added the following explanation to the text: “The mean flow has been calculated as the arithmetic mean of the daily flows within the reference period (listed in table 1), while the mean low flow is calculated as the arithmetic mean of the lowest daily flows of each year within the particular reference period.”
Furthermore we will include the reference periods used to calculate the mean / the mean low flows for each gauge in table 1.

P6l14: Which version of EOBS?

Reply: In the final set-up, we used version 13.1 as predictor in the statistical approach as well as forcing for the hydrological model in simulation mode. This is specified in the revised version of the paper.

P6l30: Is there an influence of the member size differences?

Reply: We haven’t noticed significant differences, but, to be honest, we also did not explicitly investigate on this issue (until now).

First paragraph P7: the method to bias correct the data is not explained and should be clarified. It is not clear if the authors used daily or monthly data (I suppose daily).

Reply: In order to clarify this aspect, we reformulated the section above table 2: “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). We corrected daily values of the different parameter on a monthly basis, which means each daily value of the same month is corrected by the same scaling. In future applications different bias-correction and post-processing methods will be applied and analyzed. As meteorological seasonal forecasts tend to drift …”

Table 2: E-OBS is not global, it is over Europe.

Reply: We corrected E-OBS’s spatial coverage in table 2.

Last paragraph p8 and Fig. 3: Which lead time of the forecasts? What do the boxplots and the outliers represent (inter annual variability)? R and NSE are calculated for the entire period or just for the climatology curves?

Reply: Figure 3 don’t show forecasts but the simulated flows using the hydrological model (that means the hydrological model forced with measured meteorology). The distribution of the mean monthly flows (measured and simulated) is shown as box-and-whisker plots: the box represents the 25%–75% inter-quantile range with the median as band inside the box. The whiskers represent one and a half times the inter-quantile range and values beyond are plotted as single data points. R and NSE have been calculated based on the simulated and the observed values in order to validate the performance of the hydrological model. They have been calculated for the whole period (1981 – 2015) for monthly mean values.

In order to clarify this, we extended the text concerning this matter (boxplot, calculation R and NSE).

P9l19: Sorry, it is a bit confusing, the ‘measured discharges at the forecasting gauges’ is an observation, right?

Reply: That’s correct. In order to clarify this, we switched to the term “observed” discharges in the revised paper.

P9l20: ‘based’, do you mean driven by?

Reply: That’s correct. We used the term “… forced with observed meteorology” in order to clarify this in the revised paper.

P11L7: ‘climate and hydro-meteorological variables as predictors instead of climate indices.’ Not clear, what do the authors call climate indices here.

Reply: We refer to calculated quantities / values describing a certain aspect of the climate system (e.g. North Atlantic Oscillation, Southern Oscillation Index) as climate indices. In contrast we refer to measured quantities, like sea surface temperature or precipitation, as climate and hydro-meteorological variables. In order to clarify this, we added the aforementioned examples to the text.

P12L12: A reference is missing!

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 …”

P12L14: MoNM7Q

Reply: We corrected this.

P12L25: The skill score is not necessary well defined since it tends to dump the scores when the reference is low. What are the reasons of this choice?

Reply: Yes, it’s true that the choice of the reference forecast has a significant effect on skill scores. But on the other hand the use of skill scores offers several advantages. A relevant advantage for us is that such a score could be explained / communicated quite well even to non-scientific users. We would like to proof the “added benefit / skill” of the different forecast approaches, which is supported by skill scores very well. Another aspect is that by using skill scores, different skill measures could be compared.
In our case we used climatology and ESP-based forecasts as references, which we think is justifiable:
i) Climatology represent the current procedure to estimate flow conditions on monthly and seasonal time-scales due to missing forecasts. ESP is an approach relatively simple to implement without the need of meteorological forecasts.
ii) Both benchmark forecasts don’t show extremely low skill and a more suitable in our case as e.g. forecasts based on persistence.
ii) Climatology is a typical benchmark used in various studies.

P15l18: The limitation is only over Central Europe?

Reply: Certainly not; in order to avoid misinterpretations, we slightly modified the sentence to: “Although it is well known that Central Europe is a region offering limited skill of seasonal

meteorological forecasts, in particular for precipitation as the most-important input to hydrological models, the question was if the information of these forecasts could provide some additional information to the seasonal hydrological forecasts."

Fig. 6: I am very impressed by some scores depending the months. For instance, the same skill score provided by S4 with 4-month lead time in March and 1-month lead time in August.

Reply: Yes, that is what we found out. But the results shown in figure 6 support our initial expectation that the skill along Rhine, Danube and Elbe is extremely sensitive to the initialization month. For each initialization month skill decreases with increasing lead-time, but amongst the same lead-time of different initialization months significant differences in skill became visible. The effect mainly results from the impact of the initial conditions as the S4-based and ESP-based forecast show very similar patterns. Unfortunately this relatively long-lasting skill isn't detectable in summer / autumn, when low flows typically affect waterway transport.

Figure 7: In the caption, next month or current month (1 month lead time)?

Reply: Thanks, it should be "current month" as the monthly forecasts in our case are issued at the beginning of the particular month (so 0 month lead time). It is adapted in the caption of figure 9, 10, 11 and 12, too.

Figure 8: Why there is only 4 months here? 3-month moving windows should be used.

Reply: Up to now, the 3-monthly forecast based on the statistical approach is issued just 4 times a year at the start of the meteorological seasons (e.g. meteorological spring includes March, April and May. We plan to extend also the statistical approach to generate 3-monthly forecast issued each month (like a moving window).

Figure 11: In November, S4 has less member, is it correct than June and September, is it correct?

Reply: No, the (re-) forecasts issued in November have more members (for the period 1981 – 2011) than those issued in June and September (see page 7, line 1-2).

Conclusion: No new results in this section (Tab. 3)

Reply: Due to the remarks of M. Zappa (first referee), we added a new section 5 "discussion". We decided to move table 3 and the related text to this section.