Report on *Skillful Decadal Prediction of German Bight Storm Activity*

a manuscript submitted to NHESS by
Daniel Krieger, Sebastian Brune, Patrick Pieper, Ralf Weisse1, and Johanna Baehr

May 26, 2022

1 General Remarks

In their study *Skillful Decadal Prediction of German Bight Storm Activity*, the authors aim at adressing a relevant and interesting aspect: a near term climate prediction of storm activity in the German Bight. Language and structure of the manuscript is mostly ok and it fits in the scope of the Journal. However, the description of data and methods is not sufficiently clear and the language is not sufficiently precise. Some concepts and steps in the analysis are described in a somewhat sloppy way and I expect that it will be difficult to reproduce the work described here. Furthermore, the way some methods are used does not seem to be adequate in some cases. The basis for drawing conclusions is thus not robust and the manuscript needs major improvements in several respects before it can be published. I suggest that the authors have a look on the comments and suggestions below. When ever possible, I suggested already solutions to the comments and questions I have in oder to make improving the work as easy as possible. However, there is still a lot of work involved.

2 Major comments

2.1 Conclusions

I list some of the conclusions drawn in the manuscript and comment on them.

- “Over the North Atlantic region certain regions with significant skill emerge, but the skill is dependent on the length of the averaging window.” and some sentences later: “We hypothesize that this lead time dependency might be attributable to the filtering of high-frequency variability by the lnger averaging window, in combination with the model’s ability to better predict the inderlying low-frequency oscillations in the large-scale circulation.” There is effect from estimating correlation coefficients from autocorrelated series which might – at least partially – account for this effect, see Sec. 2.4.1. The authors should take this into account for their conclusion.

- "... the DPS generates skillful probabilistic predictions for extreme low and high ..." and "As this stands in contrast to the deterministic predictability of winter MSLP anomalies, we want to emphasize that we do not have a convincing explanation for this ..." I suggest to discuss this in relation to the effect of the choice of the reference forecasts as described in Sec. 2.4.3 and the above mentioned effecte of estimating correlations from smoothed series.
- "Highly aggregated probabilistic skill scores[1], which aim at incorporating the model performance for various categories into one single value, might underestimate the capabilities to predict extrmes ..." I suggest to consider the discussion in Sec. 2.4.2 and revisit this conclusion. .

I am surprised that you have not discussed the influence of a drift or initialization shock on your forecasts.

2.2 Terminology

Throughout the manuscript the authors use the term "skill" in its colloquiual meaning as "ability to do something" and also in its special meaning within the frame of forecast verification as the value of a skill core. Also other members of the community use this unequivocal way of using the word "skill". However, I think it does lead to confusion. For example, it leads the authors to comparing "skill" of a deterministic forecast (measured with anomaly correlation, which is not a skill score but only a part of an accuracy measure) to "skill" of a probabilistic forecast (measured with the Brier skill score). This is not meaningful, see also below.

You frequently use "deterministic skill" and "probabilistic skill". However, not the skill is probabilistic or not; it is the forecast which is probabilistic or not.

2.3 Structure

The section 2.2.2 "Pressure reduction and geostrophic wind" should be renamed to "Geostrophic Wind and German Bight Storm Aktivity" or only "German Bight Storm Aktivity" as this is the goal, as far as I understand. Deriving MSLP and the geostrophic wind are only means to arrive at the GBSA, right? I suggest to show also a time series and a histogram of GBSA.

Maybe you should subdivide Sec. 2.3 into 2.3.1 "Anomaly correlation" and 2.3.2 "Brier Skill Score" to emphasize that these are two really different concepts.

The text from line 304 onwards might also fit well in to the discussion section.

2.4 Statistical concepts

2.4.1 Anomaly correlation

When introducing the correlation coefficient, you do also mention the fundamentals of your significance test, which I consider as important! However, I have some doubts that

[1] I assume that the RPS is meant.

this concept will work here as described. Is there an assumption on the distribution of the f_i and o_i for deriving confidence intervals via the z-transform? If so, is this assumption reasonably well fulfilled here?

Is the significant test a two-sided or a one-sided test? I assume a two-sided so as you mark also negative ACC as significant but it would be good to clearify.

Later, you use this correlation coefficient for measuring the association of two time series obtained from averaged (over a period of years) forecast and observations. This implies, that successive values o_i (and f_i) which enters the estimation of your correlation coefficient are not independent. This reduces the effective number of data points (degrees of freedom) and thus, you have less than N independent data points. This alters your confidence intervals (or critical values for your significance test). In Fig. 1 I give a simple example with associated code in Appendix A. You can see two effects here: i) with increasing size of the averaging window (increasing autocorrelation) the distribution of the correlation coeffcients gets broader (see Fig. 1, left) as there is less information introduced with every new paar of data as they become increasingly dependent with increasing averaging window. This effecte should be reflected in calculating

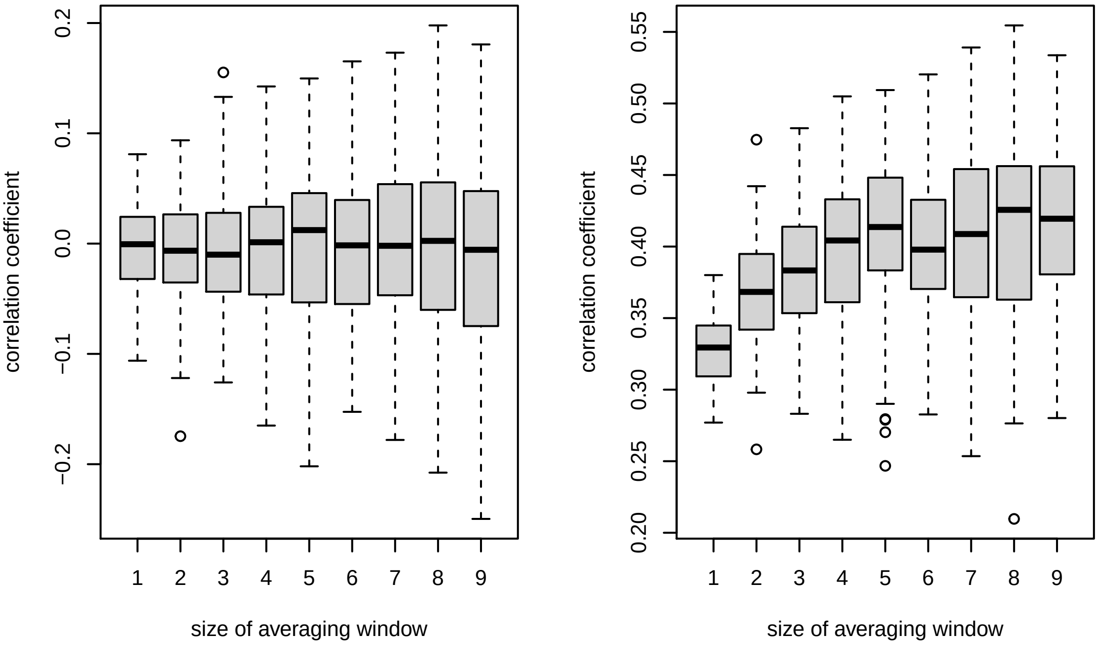

Figure 1: Example for changing distribution of the correlation coefficient with increasing size of an averaging window. Correlation coefficients are obtained for 100 series of length 1000 data points. Left with no association between the two series, right with a linear association between the series. See Appendix A for the code.

the significance of your correlation (see **?**). ii) in case there is some association between your forecast and observation, this association will seemingly become stronger for increasing averaging periods (see Fig. 1, right). This might be (at least partially) the reason for your observation that longer averaging periods lead to increased correlation (your Fig. 2).

2.4.2 Nature of the probabilistic forecast and Brier score

It is necessary to motivate the discrepancy of obtaining probabilistic forecasts for *three* categories and the use of the Brier score (evaluating *two* categories). If your aim is to issue probabilistic forecasts for the three categories you define, than this forecast can NOT be evaluated with the simple Brier score. Instead, you need the extention of the Brier score to three categories (RPS). If your aim is to issue a dichotomous forecast for the probability of being e.g. in the upper category, than the Brier score is adequate. However, then there is no need to define three categories. Currently one could have the impression that you want to issue a three-category probabilistic forecast but choose to evaluate the *simpler* dichotomous forecast (three times). Juggling three times with two balls is simpler than juggling with three balls. Simpler tasks are likely to lead to better scores. Please consider this when discussing the "neccessity to evaluate the skill for each prediction category individually" in line 310 ff.

Observations, derived values and forecasts seem to be "standardized". That would imply that using ±1 leads to consistent three categories over all the time series used. It would have been inconsistent If you instead had choosen fixed values $\mu \pm \sigma$ from the observations and use the same fixed values for the forecast. I think, it is well done here, but I suggest to clarify this explicitely. You mention in l.188 that "all" series are standardized. To me, that could refer also only to all observation and derived series.

Please also mention how the probabilistic forecasts for the categories are obtained. I guess this is from counting members with forecasts falling in one of the three categories, right? It is also important to report how you obtain the forecasts for averages of certain periods. Are the series obtained from sliding averages for longer periods also "standardized" again? Or are these categorized according to the individual lead years by their original value? This should make a difference!

It seems somewhat arbitrary that you choose the period of lead years 4-10 for your forecast. Please motivate that.

2.4.3 Reference forecast

There are a few canonical reference forecasts in meteorology, e.g. the climatological forecast, the persistence forecast and also a random forecast. When defining these for simple one-step-ahead forecasts, we can easily derive particular characteristics for some skill scores using the above mentioned references, see e.g. **?**. Mostly, we have also a good intuition on how to interpret these reference forecasts. The persistence forecasts is motivated by the fact that there is frequently persistence in the weather system and a simple forecast for tomorrows weather is the weather of today. A forecast which cannot beat this persistence referene forecast is of no value. In the case of weather forecast, we expect that the persistence forecast outperforms the climatological forecast because of the persistence of weather systems. If we use todays weather to issue persistence forecast with increasing lead time, it seems plausible that the this persistence forecast performs worse with increasing lead time because the basic assumption of persistent weather is less likely to hold for increasing lead time. Looking at your Fig. 4, this effect might play a role when you identify increasing BSS (reference persistence forecast) with lead time. You should discuss this effect when revisiting your conclusions.

For a non-probabilistic forecast for a quantity X with probability distribution $F(x)$, a random foecast can be generated by drawing a random number from $F(x)$. This is a different value for each time you issue this forecast. For a dichotomous probabilistic forecast, a random forecast $f \in [0, 1]$ could be a uniformly distributed random number on $[0, 1]$. I would not call forecast with a fixed value at e.g. $f = 0.5$ a random probabilistic forecast. Fixed probability forecasts are common for the climatological forecast. Here your forecast is the climatological occurrence probability $f = \hat{p}$. I find the interpreation of your "random forecast" difficult. In particular when saying (l.298) that "random guessing" is performing better than "persistence". That indicates that there is something wrong with the common interpretation of either the "random" or the "persistence" forecast, or probably with both. I interpret your "random forecast" as a climatological forecast with a non-adequate estimate of the climatological occurrence probability.

To be more consistent, you could use the same reference forecasts for the non-probabilistic and the probabilistic forecasts. I suggest that the climatology would be a good reference. The MSE might also be adequate for a scoring function for your non-probabilistic forecast.

2.5 Associated to lines in the manuscript

l.9 "skill of the probabilistic predictions for high and low storm activity notably exceeds the deterministic skill" This is a comparison by numbers of two different measures of forecast quality for two forecasts of very different nature (non-probabilistic and probabilistic). Such a comparison is not meaningful. It is like saying 0.6 Kg cheese cake outperforms 0.5l water because 0.6 is larger than 0.5.

l.13 "may greatly benefit" can you give an example how planning and management can benefit from predictions on the decadal time scale?

l.32 "While Kruschke ..." As I understand the situation, Kruschke et al. evaluated a 3-category probabilistic forecast using the ranked probability score. Here, you suggest to evaluate a dichotomous (2-category) probabilistic forecast. The task you choose to address is a lot simpler than the task addressed in Kruschke et al. It is easier to forecast probabilities for 2 categories as this forecast contains less information. If you repeat this task 3 times, it remains an easier task (see my juggling example above). Both strategies (issuing a 2-category and a 3-category forecast) are valid. Which one to chose depends on the goals of your study. Please explain why you define 3 categories but chose to evaluate a 2-categories forecast. Please note that even if you devide your forecast into three categories, you evaluate with the Brier score only the ability of the model to forecast the probability for being in one categorie or in one of the other two.

l.62 “The concept of a probabilisitic approach is the presumption that a shift in the ensemble distribution can be used to predict the likeihoods of actual shifts in climatic variables.” Please clarify what you mean with this sentence. From my point of view, the concept of probabilistic predictions is to include the uncertainty in the forecast while this is completely ignored for non-probabilistic prediction.

l.74 “... skill is quantified by correlating ...” The correlation coefficient measures the linear association between two series. It is part of the MSE (a score masuring accuracy) and also part of the associated MSE skill score (measuring skill with respect to the climatological forecast). Only in special cases the MSE skill score with the climatological forecast as a reference is proportional to the correlation coefficient between forecast and observation, see **?** or **?**.

l.81 Explain what you mean with “standardized 95th percentile”. Later in the manuscript you describe “standardization” as the z-transformation $z = (x - \bar{x})/sd(x)$. I assume that this is what you mean by “standardization”. Please describe here what you do, why you do it and why this is adequate. It might help to show a histogram of wind speeds and/or the 0.95 quantile of wind speeds.

l.146 Here, I suggest to that you describe in some detail HOW you generate the time series. Do you use the the gradient from a plane through 3 grid points as in Krueger et al? Or is there some averaging involved as you mentioned in l.54?

l.147 Explain the standardized annual 0.95 quantile and show a time series and histogram plot.

l.154 Maybe you want to change the title to “Evaluation of forecast/model performance”

l.155 In fact, you obtain a skill measure (Brier skill score) and a measure of linear association, which I would not call “skill”, see Sec. 2.2.

l.166 Please include in this sentence that you can evaluate *only* dichotomous probabilisitc predictions (i.e. probabilistic predictions for two categories) with the Brier score. The natural extention of the Brier score for more categories is the RPS and for continuous forecasts the CRPS.

l.173 briefly explain the bootstrap approach used here. What is sampled? With replacement?

[l.174] “... test whether skill scores are significantly different from the reference.” → either say “test whether skill scores are significantly different from zero.” or “test whether model performance is significantly different from the reference”.

l.176 "The individual Brier Scores BS are defined..." → "The Brier Score BS is defined as ..." (In fact, the equation you give is already the average Brier Score over N forecasts.

l.179 ".. the predicted event .." happened. → "... the event happened". The observation can be 1 even if you have not predicted the event.

l.180 This is a probabilistic prediction, it can be better or worse but not "correct" or "wrong". Change the sentence accordingly using the range of the BS and its negative orientation.

l.182 The "random guessing" does not lead to $f = 0.5$, see Sec. 2.4.3.

l.185 if all your series are "standardized" the $\mu \pm \sigma$ interval should be ± 1. If not, you need to give the μ as well.

l.189 GBSA is derived from spatially averaged MSLP gradients? From your description, I thought it is derived from a gradient of the plane through three grid points. Clarification needed.

l.188 is GBSA also standardized? Clarify.

l.194 "... persistence prediction of storm activity is generated by taking an average storm activity .." but this must be a probabilistic forecast between 0 and 1, not a storm activity. Clarify.

l.192 ".. we use a persistence prediction.." it is rather a "kind of persistence". Murphy used it in a clearly defined way. In your case, the quality of the "persistence" reference depends on lead time, see Sec.2.4.3. You might want to check the BS of the persistence as a function of lead time (plot).

l.204 please be more precise. Positiv anomaly correlation is not equal to a positive skill score.

l.205 Here you relate negative (and significant) correlation (from your Fig. 1(a)??) to "significant skill". Please clarify.

l.209 same here.

l.210 "... overall magnitude of the ACC is lower .." please be more precise. Do you mean that absolute values are on average lager?

l.211 is that statement not a repition from l.208?

l.212 please relate this finding to my remark in Sec. 2.4.1.

l.262 please discuss the performance of the reference prediction (persistence), ideally with a plot depending on lead time.

l.263 please reconsider the comparison with the "random prediction" under consideration of my remark in Sec. 2.4.3.

l.265 please define what you mean with "absolute skill"

l.266 reconsider the comparison to "random guessing"

l.267 it is correct (and trivial) that the "extreme anomaly events" (in your case those above $\mu + 1\sigma$) occur on average at a rate of $(1 - 0.68)/2 = 0.16 < 0.5$.

l.297 I agree with the low performance of the reference and suggest that you use climatology as a reference. That would challenge the model more than your "random guessing"

l.311 please reconsider this paragraph with respect to the discussion in Sec. 2.4.2.

l.316 maybe a plot of the BS for the reference helps to understand that.

l.323 "random guessing is ill-suited" and "therefore persistence ranges among the most appropriate references", see Sec. 2.4.3. I suggest climatology.

l.325 "Our DPS is particularly valuable at lead times during which persistence forecasts are sufficiently poor". That does relate to the discussion of the interpretability of the persistence forecast in Sec. 2.4.3. I would rather discuss this as a deficit in the choice of the reference forecast and *not* as a "valuable" aspect of the DPS.

l.335 I am not sure how relevant the assumption of normal distributed quantities is. Certainly, you only get a standard normal distribution if your variables are sufficiently normal before standardization. You can get rid of this assumption if you use the 0.16 and the 0.84 quantile of the distribution to categorize your events instead of $\mu \pm \sigma$.

3 Minor comments

l.9 "short lead years" → "short aggregation periods"

l.10 "deterministic skill" not the "skill" (or performance) is deterministic, but the forecast. → "performance of the deterministic forecast". But I recomment to *not* compare these two anyway, see above.

l.11 "... skillfull decadal predictions of regional storm activity can be viable .." → "decadal predictions of ... can be skillfull"

l.16 "... attributed to the anthropogenically caused global warming trend ..." → "... attributed to the anthropogenic global warming ..." (Okkhams Razor).

l.23 "... great potential value in moving ..." → "... great potential in moving ..."

l.23 ".. for each separate category." → "for each category"

l.53 "..., since averaging over a small area preserves much of the spatial variability .." please specify what you are doing here. Spatial averaging does not preserve spatial variability. The opposite is the case.

l.56 "... besides the choice of parameters ..." → "besides the choice of variables"

l.62 "... benefits the generation of probabilistic predictions." → "is favourable for probabilistic predictions."

l.64 "... of the prediction distribution ..." → "predictive distribution"

l.65 "viable" → "feasible"

l.73 and in other places "... deterministic/probabilistic skill ..." → "skill/performance of a deterministic/probabilistic forecast"

l.86 ".. derived observational time series" What do you mean by that? If you derive a quantity from other quantities, this cannot be an observed series. Do you mean, that the series under question been derived from observations?

l.109 "constrict" → "constrain"

l.251 "lead year ranges" → "averaging periods"

l.286 what is the "ensemble BSS"

l.290 What is a "skill difference". Skill is already the difference of an accuracy score of a model and a reference (scaled by the difference of the perfect forecast and the reference).

l.291 "most valuable at skillfully predicting" → "valuable at predicting"

l.292 "which demonstrated significantly skill" → "which significantly demonstrated positiv skill"

l.302 ".. not viable to skillfully predict .." → "... is not skillful in predicting ..."

l.319 "overwhelmingly significant" does not exist. A result is either significant on a certain level or it is not. Less or more significant does not exist. But you can give significance on various levels.

l.351 "lead range length" → "averaging period"

l.352 "non-single lead times" → "averaging periods larger 1 year"

A Simulation experiment for correlation of smoothed time series

The following code gives a simple example written in R of how the distribution of the correlation coefficient might change if the two series are smoothed with a running window. In the example, I use a simple autoregressive process to mimic the observations and the forecast.

```
library(zoo)
N <- 1000
M <- 100
a <- 0.3
mas <- c(1,2,3,4,5,6,7,8,9)
cors  <- cors.2 <- matrix(NA,ncol=length(as),nrow=M)
names(cors) <- as

for(i in 1:length(mas)){
    for(j in 1:M){
        o <- arima.sim(list(ar=a),N)
        f <- arima.sim(list(ar=a),N)
        f.2 <- o+rnorm(N,sd=3)
        cors[j,i] <- cor(rollapply(o,mas[i],mean),rollapply(f,mas[i],mean))
        cors.2[j,i] <- cor(rollapply(o,mas[i],mean),rollapply(f.2,mas[i],mean))
    }
}
layout(matrix(1:2,ncol=2))
boxplot(cors,xlab="size of averaging window",ylab="correlation coefficient")
boxplot(cors.2,xlab="size of averaging window",ylab="correlation coefficient")
layout(1)
```