Reviewer(s)' Comments to Author:

Reviewer: 1

Comments to the Author

Simonis et al.: Evaluating probabilistic ecological forecasts

This study exemplifies the evaluation of near-term time-series predictions (“forecast”), discussing the internal validation setup and the scores used to measure forecasting discrepancy. While the paper is proposing a best practice, I found several aspects not well covered. That said, the paper is in its gist and spin extremely important and the ecological community is increasingly employing improperly evaluated forecasts, making this contribution very timely. It is because I like it that I write rather a few comments.

1. Unbalanced treatment of all steps

I missed an overview of all the issues that need to be addressed when evaluating probabilistic time-series forecasts. Maybe a step-by-step guide would be useful for the reader, and then the authors can choose which to exemplify and which to “only” reference.

1 decide on a cross-validation approach

2 fit a model to a subset of the data

3 analyse the temporal autocorrelation of the residuals of this model

4 generate predictions including prediction uncertainties for the hold-out

5 evaluate the observed hold-out on the prediction distribution

6 repeat from 2, depending on you choice of 1; possibly rethink 1

Ad 1) Now, in the present paper, the authors do not discuss cross-validation options. They instead refer to other papers. I’d like to draw their attention to a paper on the arXiv, which I regard as a very nice review of cross-validation options for time-series:

@article{Cerqueira2019, title={Evaluating time series forecasting models: An empirical study on performance estimation methods}, volume={abs/1905.11744},, journal={arXiv}, author={Cerqueira, Vítor and Torgo, Luís and Mozetic, Igor}, year={2019} }

Among other things, this paper makes a good argument for NOT following the route taken by the authors of the present paper: if temporal autocorrelation ranges are long, we need to put some distance between the end point of the training data (what the authors call o) and the first prediction (what the authors call p). Such a buffer prevents non-model prediction to spill over into the evaluation set. This is in line with the review of Roberts et al. (2017 Ecography) on cross-validation for non-independent data (dear authors: sorry for the self-advertisement: you don’t need to cite it, only absorb the message). The rolling evaluation employed in the present study is, in my reading, indeed recommended.

I strongly urge the authors to communicate clearly which different cross-validation options exist (possibly in a table akin to that for the scoring rules), and that we need to balance the bias introduced by temporal autocorrelation against the loss of skill when training of ever-shorter time-series.

Ad 2) That is largely standard time-series statistics, and the Bayesian approach employed by the authors is great, although for the “frequentist” reader an unnecessary distraction, I find.

Ad 3) The authors fail to describe the model residuals and the remaining temporal pattern. This is IMHO crucial for deciding on step 1, the type of cross-validation that is valid.

Ad 4) To me it did not become clear, how prediction distribution H is produced. Now, I would have a way to do that, and I guess the authors’ approach in JAGS is probably the same, but this should be spelled out more explicitly, and particularly for non-Bayesian models. Say somebody fits a GAM with an AR1 term, which I would consider a sound default approach, then this GAM’s prediction to the hold-out will also include a standard error for the model prediction. However, the standard error introduced through the AR1 is (AFAIK) NOT included in this prediction. Thus, the SE is only valid for data beyond the autocorrelation range of the model residuals (it is smaller within the range). I am not saying this is how it should be done, my point is that the authors must make clear, how the prediction distribution H is generated, which currently they do not.

Ad 5) I have to shamefully admit that I did not know the PIT (under this name), although I have repeatedly used it and always referenced one of Andrew Gelman’s posterior p-value paper. I now lost the point where the scoring rules come in. Is the PIT applied to the scoring rule scores, rather than the prediction distribution H directly? If so, why? I guess I lost the plot here, and maybe I am representative for the interested reader.

Ad 6) Fine, no comment.

2. Discussion of discrepancy between within-sample and out-of-sample PITs.

This is a pattern I have often observed myself when applying this (or a similar) scheme: the model evaluations look more or less great within the rolling evaluation, but the last few data points are really different. I think that is, to a large extent, a bias introduced by the analyst. If the environmental data do not change, why bother analysing them? So we really turn our attention to data when something does change. And such a change is often caused by changes to the system itself, say climate change, habitat loss, eutrophication, you name it. This would mean that in the last few time steps, the system underwent a change, the time series is no longer stationary and hence the uncertainty so far is no good predictor for the uncertainty in the future, even if we get the mean more or less right (which we might not). I do not know whether this interpretation applies to the pocket mice in the example here, though. Still, I think the point that any extrapolation (“forecast”) assumes stationarity (in the wider sense), i.e. assumes the system to not change in any fundamental way, deserves a mentioning.

Details (I switch to directly addressing the authors for ease of writing):

\* Abstract: Please do mention methods and recommendations. This is a summary, not an appetiser.

\* Intro: Define, even vaguely, what you mean by “forecast”. “Short-term extrapolation of a time series” is one option, although “short-term” means something very different for a soil microbe than for a blue whale.

\* L30: add something like “as found by” into the brackets, otherwise the reader doesn’t know whether these references are just examples of bad practice.

\* L43: add “(mathematically: the population)” after “distribution”. I understand that you don’t want to call G the population for fear of confusing the statistical and the ecological term. But I find it curious to use the fashionable “data-generating model” lingo without reference to the perfectly acceptable sample/population terminology. If it is unsuitable, please let the reader know why.

\* L46: Not necessarily o+1! As I wrote above, it may well be o+buffer+1 or so.

\* L47: end of sentence: repetition from before.

\* L48: subscripts of y: (o+p):(N+p+P), where p is the buffer. Here would be a good place to add the issue of having a buffer at all, depending on the range of temporal autocorrelation.

\* L49: both refers to what: fit and forecast?

\* L50: why single out prequential (see comment to 1 above)?

\* L57: Delete “the dominant … validation”, and “, rather than … (Stone 1977).” Here is a good place also to refer to the other approaches as reviewed by Cerqueira et al.

\* L79: I can see no reason why the train data should have a minimal length relative to the prediction data. If the model is suitable, the errors will of course be much larger, but not biased. You pass the “should” on from Tashman without stating for which purpose. For acceptable errors or bias? What is acceptable, for whom? That is not for you or me to decide but for the analyst. So I dislike the “should” in methods, unless clearly justified. It may be here, but I doubt it. Rather, it is a rule of thumb, which is fair enough for ecologists, but you plug it into a very technical part of a paper, where it is out-of-place.

\* L81: Same.

\* L84: I don’t think cyclic dynamics are at all common in ecology. Prominent in Nature papers, but not common at all. The analyses of the Global Population Dynamics Database do not show more than a handful of examples.

\* L84: Why should cyclic dynamics cause the origin of the forecast to matter? The model should represent the cyclic behaviour, and if it didn’t, the variability over time would lead to very uncertain but IMHO not to variably uncertain predictions. There may be a reason which I overlooked, but then please argue for it.

\* L91: I don’t understand: who would NOT update the model? Of course you would have to recompute it (or Bayesian-like update it) to the expanding training data. Please do not open an option that would be ignoring data and which you also do not choose for good reasons.

\* L93: “may not” or indeed “may dramatically”. I see no reason why to encourage poor statistical analysis by suggesting NOT to update. If there is no change in parameters, all that is lost is a bit of computing time. But if there is a change in parameters, then it would bias the model prediction NOT to recompute the model.

\* L94: What is “recurrent” here? It is not in the cited references (Dawid and DietzE, with an e in the end). Maybe rephrase these few sentences. In deep learning, recurrent does not mean updating, so it may be misinterpreted. I like updating (and a bit less: online learning).

\* L103: Why should that be important? I want to see the prediction and its error, not how the model fits to the past. If that would be informative, then the model is probably poor! Of course the analyst should plot this, but hope for this plot to be ideally Uninformative.

\* L105: Why should the predicted not follow the 1:1 line perfectly? What you mean is that the prediction should not have too small error, so that the precision is incorrect (optimistic). But this way of phrasing is misleading.

\* L119: Maybe add Gelman’s paper on Bayesian/posterior p-values, broadening the range of people that can connect with the PIT.

\* L123: “T”able 1

\* L142: “observations are binarily matched or not”: I don’t understand a word. What does that mean? Binary what matched or not how?

\* L146: Which variance is “inflated” by autocorrelation? That of the prediction (no, it should be deflated)? That of model parameter estimates? My point is that “inflation” is often the correct statistical quantification of model uncertainty, and thus a good thing (even if it leads to larger error bars).

\* L176: What is the “marginal predictive distribution”? I know all three words, and all combinations of two of them, but not this triplet. As it is a reference for the skill score, please do not only refer to another paper.

\* L182: What do you mean by “correlation among values”? Is that the same as temporal autocorrelation?

\* L185: Same here again.

\* L186: differential –> difference (differential suggests a derivative, not a difference, but a plain old difference it is, according to the appendix). And: Difference of which 2 scores or which 2 forecasts?

\* L188: “difference” again

\* L189: Is “serial” autocorrelation the same as “temporal” autocorrelation? If not, then what? More important: this is surprising to read, since you first tout this test to account for correlated errors, and now you say, actually it needs to be addressed with robust formulae. This is strange and inconsistent. So what does and does not the D-M-test do?

\* L197: add “right-” before “truncated”

\* L200: this is not cycling, this is seasonality: 12 months make a year, you know.

\* L213: The PIT-diagramm peak does, in itself, NOT indicate bias. By the looks of it, the average value is probably very close to 0.5 (if the x-axis range is 0-1). I would thus disagree that there is a bias, only that the coverage of the prediction is too wide.

\* L214: Here you should point out that the out-of-sample PITs are worse than the in-sample PITs, which I guess is fairly typical.

\* L217: This statement is about a case study, so I would appreciate if you add an “in this case”. I cannot remember where I read about many different time-series structures, but AR1 is a fairly well-performing model structure overall. However, this is no proof of it, only a demonstration. Again, remove “biased”.

The discussion is boring and does not make any point related to the study. It first repeats the introduction, then writes general stuff about statistics in ecology. I would not mind to see it go.

Instead, I think it would be useful to offer some recommendations, tentative and carefully worded, about how to carry out the six steps I listed above (or 5 or 8). Personally, I would also like to see the key references for each, should I like to read up on it.

Also here you could add something along my pet lines of arguments: When someone takes a decision, based on a wrongly quantified uncertainty of a forecast, the forecaster is to blame! Thus, we should aim to compute correct prediction distribution H, not only correct point estimates. Otherwise the decision maker may think: “Ah, it’s quite certain that we will get a value of 5! Let’s act on that.” when in fact it is “5 +/ 4” and that may require a much more guarded action. Just a suggestion in order to encourage you to strongly justify why we should spend time on getting the entire prediction distribution right.

Or you could talk about using the rolling hold-out validation as a quantification of the actual prediction uncertainty, rather than the prediction distribution provided by the model itself.

Table 1: I am confused about the notation for the quadratic score. Now, f(y\_n) is a vector, but ||f(y\_n)||\_2 is NOT (as defined in the table footnotes, it is a scalar, as it sums over y). This is not correct, I think. Or the sum is over something not explained but not over the data. In wikipedia, it sums over “classes”, and this is probably the same issue that I had with the “binary matching” earlier. So please clarify at least the index of summation.

Fig. 2: What are the values of the x-axes in panel b and c? Could you please also provide the mean value here (as Gelman recommends for Bayesian p-values, aiming for a 0.5)?

In the caption you could add “monthly” before the species name.

(d) is too small, and the overlap of violins is so tight, that I cannot distinguish them. I suggest moving the three models out a bit, next to each other (still overlapping).

(b) should have a light grey background

(c) should have a dark grey background

b and c should have ticks and labels on both axes.

Carsten Dormann