Reply on first review

Calibration of medium-range metocean forecasts for the North Sea

by Murphy et al

Please address all correspondence to Philip Jonathan (*p.jonathan@lancaster.ac.uk*)

Summary

We thank two reviewers sincerely for their time and expertise in reviewing the manuscript. We note that both reviewers find merit in the work.

Reviewer 1 is more-or-less happy with the original form of the article. Reviewer 2 would like to see direct out-of-sample evaluation of the generalisation error of the different models, and suggests using CPRS as an additional measure of model performance. The major change to the manuscript, in response to Reviewer 2, is inclusion of an evaluation of generalisation performance of previously fitted models using more recent data. We have also included plots of CPRS.

More generally, both reviewers make suggestions for improvement of the manuscript, which we reply to below, entailing some additions and modification of the manuscript. Where we feel the manuscript is already adequate on a point, we provide a note of explanation.

In the following sections, reviewer comments on the manuscript are given verbatim in italics and referenced by 'R'. Our replies (prefixed 'C') are given in normal font per review comment. Larger changes to the manuscript are given in red in both rejoinder and revised manuscript. References to locations in the article are made by us per section (e.g Section 4, S3.5), to avoid confusion with page- and line-number differences between manuscripts.

We hope that the revised manuscript is acceptable for publication, but would be happy of course to revise further in light of additional review comments.

Reviewer 1

R1.0 This paper addresses calibration of weather forecast for offshore applications. This is an interesting topic of relevance to the journal. The paper is well written and the analysis appears to be carefully carried out and reported. Hence, I can recommend this paper for publication in Applied Ocean Research.

C1.0 Thanks.

However, some minor points the authors want to consider in preparing a revised, final manuscript are given in the following:

R1.1 Figures are difficult to see even in colour, and using the same plotting symbols makes it impossible to distinguish what is what in b/w (I assume printed version will be in B/W). (For example, the red circles mentioned in Figure 8 can hardly be seen). I would suggest that you redo the figures and use discernable plotting symbols; rather than just discs in different colours, consider to use different plotting symbols in different colours (e.g. disc, square, triangle, star, ...)

C1.1 Colours were chosen specifically so that the different figure components would be distinguishable in black and white; some of the authors have used this colour scheme successfully in previous publications. However, since both R1 and R2 have raised this point, we have attempted to change the colour scheme slightly, hopefully to provide better distinction in colour, whilst maintaining good black and white characteristics. We have also tried to use different symbols within reason to clarify plots.

***ROSS: do it (-;)

- Play with colours
- Play with symbols

R1.2 Are the forecast of the different quantities (wave height, wind speed, wave period) independent, or are they output from the same model, and could hence be considered joint forecasts? Discuss the implications of this in using them as covariates for the other quantities.

C1.2 Thanks. We have already noted in the discussion that there may be an opportunity to exploit the joint structure of metocean variables to improve forecast calibration performance. We are aware of the existence of a huge literature across many fields on multivariate regression and calibration. We would note however that such approaches are likely to be beyond the capabilities of the typical metocean engineer, and therefore less attractive in the context of the current work.

We have modified the text in S5 to read We might also consider joint calibration of multiple metocean variables. Multivariate predictive modelling is a large field of research and applications, offering a wealth of modelling strategies to predict a multivariate response from multivariate predictors. The presence of correlated predictors can lead to inflation in estimated model parameters (which can be quantified using measures such as the variance inflation factor), and inflation of prediction uncertainty. Nevertheless, joint modelling of multiple metocean variables provides the potential for better calibrated forecasts, including extremes. In the context of weather forecasting, Allen et al. (2024) provides a discussion of methods for assessing the calibration of multivariate probabilistic forecasts. Extension to calibration for multiple locations is also possible; for some locations at least, ...

R1.3 Fig. 11 is interesting and illustrates the effect of the calibration. Most notably, it is clear that the deterministic forecast of Tm is biased. However, even though these biases might be statistically significant, I think a discussion on the practical significance of this could be interesting. Will this have any practical implications? E.g. a bias for significant wave height of less than 1 cm seems negligible. Also, what is the practical significance of a bias of 0.1 - 0.15 sec in wave period?? Does it matter? This could be discussed also in light of the measurement uncertainty. What are the measurement uncertainty of Hs, W and Tm? Can you even distinguish between sea states with 1 cm and 0.1 s difference in Hs and Tm? Please consider adding a discussion on this.

C1.3 Thanks. We agree that the bias of calibrated forecasts for the three metocean variables shown in F11 is small, certainly relative to the bias in uncalibrated ensemble forecasts for the mean wave period (F2, F3) of ≈ 1 s. Of more material impact are the standard deviations of forecasts in F11 for forecast horizons in excess of ≈ 72 hours, for which 95% prediction intervals are at least ± 1 m for H_S, ± 5 m/s for W and ± 1 s for T_M. For H_S and W, these extents of uncertainties are likely to lead to more problematic decision making. We have added the following text at the end of S4. Biases in forecasts from F11 are unlikely to be of material concern to the metocean engineer, and may well be at the level of measurement error in practice. However, uncertainties in forecasts grow to levels which are likely to be of practical concern for horizons of 3 days and longer, for which 95% uncertainty bands are at least ± 1 m for H_S, ± 5 m/s for W and ± 1 s for T_M.

R1.4 Please check eq. (3). Should it be μ_{Z_k} rather than μ_k?

C1.4 Thanks for spotting this. Now corrected.

Reviewer 2

R2.0 This study investigates the calibration of forecasts of significant wave height, mean spectral wave period and wind speed for a location in the central North Sea. Linear regression and non-homogeneous Gaussian regression (otherwise known as EMOS) are adopted as simple methods for this benchmarking study. The models allow a mixture of deterministic forecasts, control runs and the mean and spread from exchangeable ensemble members as predictors, including cross-prediction, e.g. using Hs to predict W. It is found that the raw forecasts can have some biases and that forecast post-processing improves the forecast quality. I find that the paper is generally well written and is potentially very interesting to readers as an application of forecast post-processing to metocean forecasts. There seems to be a solid grasp of the forecasting methods, however, the forecast verification component is somewhat lacking. Unfortunately, I find it difficult to recommend publication until a more in-depth analysis is undertaken. I am therefore making major comments that will hopefully lead to eventual publication of this work.

C2.0 Thanks for your comments. We have attempted to modify the paper to reflect your feedback.

R2.1 Comments from pro-forma

R2.1.1 S3.3 Suggest adding detail about how NGR/EMOS parameters are estimated.

C2.1.1 Thanks. Maximum likelihood estimation is used to estimate the parameters of the NHGR model. We have added the sentence As for LR, NHGR parameters a, b, c,d and e are estimated using maximum likelihood estimation.

R2.1.2 F1 and F2. I think instead of saying correspondence is"good", some statistics like correlation could be useful to understand raw *skill.*

C2.1.2 Thanks. We believe that quantification of performance (in terms of first two moments) is provided by F3. We also feel that visual comparison in F1 and F2 provides a useful complement. We have changed the text "elaborates on these findings, by" to quantifies these finding, by, hopefully to make this clearer.

R2.1.3 Figure colour choices can be improved. It is sometimes difficult to tell apart yellow and orange (e.g. Fig 6) or to see low contrast such as yellow on white background.

C2.1.3 Thanks. See our reply to R1.1.

R2.1.4.1 It is difficult to draw any conclusions from the current study about performance of the calibrated forecasts because none of the training or validation appears to be done within a cross-validation or split-sample framework. At the very least, the training and validation strategy does not appear to be described in sufficient detail. It is critical to evaluate the performance of forecasts out-of-sample with respect to the training period.

C2.1.4.1 Thanks for this comment. We agree that the acid test for any predictive model is direct estimation of its generalisation error. However, we would point out in this work that the estimated LR and NHGR models for a given forecast horizon have at most 6 parameters, and have been estimated using hourly data over 16 months ($\approx 12,000$ observations). Even allowing for serial correlation, we can be confident that model parameters and predictive performance have been estimated reasonably, (unless of course the characteristics of future time periods are different to those of the period over which the models were estimated). As explained in the paper, model choice is made by estimating AIC for competing models. Specifically, we can be reasonably confident that the assessment of predictive performance illustrated in F11 is reasonable.

Nevertheless, to address the reviewer's concern, we have taken the models illustrated underpinning F11, and used for out-of-sample forecasting for the period ***1 October 2023 - 31 March 2024***. Out-of-sample forecast performance is then summarised in Figure SM4 of the updated Supplementary Material. From the figure, we see that ***

We have added the following text in S4.5 Provided that data for the calibration model training period is representative of the future environment, we can be confident that future model performance will be similar to that reported in F11. Nevertheless, we can also directly evaluate forecast performance for a time period following that used to estimate the calibration models. FSM4 illustrates forecast performance of calibration models estimated on the period *** for the period ***1 October 2023 - 31 March 2024***. From the figure we see that ***.

***ROSS

- Get forecast and measured data from more recent period
- Apply our existing models to this "out of sample" period
- Show that performance is ok
- New figure is SM4
- Ross, I don't think there's any point doing CPRS; it's just another score. Doesn't really tell us anything. I'll just add a reference.

R2.1.4.2 Many of the predictors are highly correlated with each other. This can lead to problems with predictive ability for future events. The authors need to justify why the ensemble mean, control and deterministic forecast should be used simultaneously and how including highly correlated predictors affects factors like ensemble spread. I think we also need to understand the sensitivity of the consistent models, identified in T1 and T2, to cross-validation.

C2.1.4.2 Thanks. See our response C1.2 regarding potential variance inflation from adoption of potentially collinear predictors in multiple regression; figures such as F7 (showing parameter estimates and uncertainties for standardised covariates) illustrate that, for the current work, there is little cause for concern. See our response C2.1.4.1 regarding the large sample used for estimation of calibration models; we would not therefore expect large variability in parameter estimates were we to employ a leave-out scheme for parameter estimation. See response C2.1.4.1 also for a direct evaluation of generalisation performance for the (calibrated) forecast models.

R2.1.4.3 F11 evaluates bias and the standard deviation of errors as a function of horizon. I would suggest that the forecasts are additionally evaluated in terms forecast reliability and accuracy. For example, the deterministic and ensemble forecasts can be evaluated using the CRPS. Ensembles can be evaluated using rank histograms/PITs.

C2.1.4.3 See comment C2.1.6.2 regarding our motivation and the "gap" we're trying to fill with the paper. We agree that there are many scoring rules that could be used additionally to evaluate forecasts, but feel that bias and standard deviation (RMSE) are adequate for the current work. We have added the following text in S5: Bias and variance are fundamental quantities used to characterise the performance of an estimator, favoured by us in the current work. Similarly, the KS statistic is a generic measure of the dissimilarity between distributions. However there are additional performance scoring rules and diagnostics, particularly interesting when evaluating ensemble forecasts, which could also be used for the current work. These include the continuous ranked probability score (CRPS; e.g. Gneiting et al. 2005) and the probability integral transform (PIT) histogram (e.g. Dawid 1984). More generally, the work of Hernandez et al. (2018) provides a review of performance, skill and accuracy assessment in operational oceanography, and Messner et al. (2020) reviews forecast verification tools, with a focus on wind power applications.

R2.1.5 There is some brief discussion of limitations in the discussion, such as under-prediction of large values. The limitations could be more comprehensively explored.

C2.1.5 We are very happy the reviewer makes this comment! Thanks. See comment C2.1.6.2 regarding the scope of the work. Under-prediction of extreme values is possibly an example of "regression attenuation" typically seen when covariates are observed with error; we also refer the reviewer to Towe et al. (2021), where a simple calibration model is applied for optimal unmanning of an offshore structure. More generally, all of the authors come from an extreme value analysis background: the calibration model involving extremes of one or more of forecast and measured values probably should not take the same form as that for the bulk of the distribution. Instead, a joint or conditional extremes model (see e.g. Heffernan and Tawn 2004, Jonathan et al. 2013, Towe et al. 2023a, Towe et al. 2024, Murphy et al. 2024) might be appropriate. There are opportunities also to exploit the extreme time-series structure of predictors and response, perhaps following from the model forms of Winter and Tawn (2016), Tendijck et al. (2019) and Tendijck et al. (2024). These are areas of active academic research, and cannot be addressed adequately in this work. Nevertheless, we are excited to try to contribute here! We have modified the text of S5 to read: The choices of calibration models used here represent the simplest approaches that might reasonably be adopted in practice. Consequently there are many opportunities to extend the analysis. We noted evidence in the exploratory analysis that the forecast model generally tends to underestimate the very largest values. This might be an opportunity e.g. to include quadratic and higher order terms in covariates in the parametric form for the forecast mean (and, within the NHGR framework, for the forecast standard deviation). Alternatively, given that joint largest values of measured and forecast variables can be considered extreme, it might be more appropriate to adopt extreme value models (e.g. Davison and Smith 1990, Heffernan and Tawn 2004, Jonathan et al. 2014, Towe et al. 2023b, Towe et al. 2024) to characterise these regions more correctly, or more generally to relax the assumption of Gaussianity made by both LR and NHGR. There are opportunities also to exploit the extreme time-series structure of predictors and responses, following the Markov extremal model and related frameworks in Winter and Tawn (2016), Tendijck et al. (2019) and Tendijck et al. (2024).

R2.1.6.1 S4.1 and S4.2 could be in the methods rather than the results.

C2.1.6.1 Thanks. We like this suggestion, and have moved the two subsections, modifying the introductory texts of S3 and S4 in the appropriate manner, mainly by moving the text The performance of calibration models with different levels of complexity is evaluated as a function of forecast horizon by comparison of values for the Akaike Information Criterion (AIC), as explained in Section 3.4. Further, for ease of interpretation of regression output, we choose to standardise the covariates in a particular manner, as discussed in Section 3.5 from S4 to S3.

R2.1.6.2 The introduction currently reads like a review of forecast calibration methods. While many of these studies will be useful to be cited, I suggest focusing the introduction on positioning the current study and the research gaps.

C2.1.6.2 Thanks. As noted already in a number of locations in the original submission, we write the paper from the perspective of a practicing metocean engineer who typically uses the simplest approaches to calibration. The purpose of the brief review of forecast calibration is to emphasise that forecast calibration a big topic (as stated in the original text), and perhaps to encourage the reader to explore some of this literature further.

We have added the following text in the introduction to clarify "the gap" that we are attempting to fill with the current work. The typical practising metocean engineer uses the simplest tools (e.g. linear regression) for calibration, rarely exploiting the richness of data provided my modern forecast models. The aspiration of the current paper is to demonstrate that there is material benefit from exploiting the output of modern forecasts more fully within pragmatic forecast calibration procedures for realistic offshore application. Specifically we do not claim that the calibration procedures considered are the best currently available, but we do contend that they provide useful tools with which the metocean engineer can reasonably expect to improve the performance of their forecast calibrations.

References

Allen, S., Ziegel, J., Ginsbourger, D., 2024. Assessing the calibration of multivariate probabilistic forecasts. Q. J. R. Meteorol. Soc. 150, 1315–1335.

Davison, A., Smith, R.L., 1990. Models for exceedances over high thresholds. J. Roy. Statist. Soc. B 52, 393.

Dawid, A.P., 1984. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. J. Roy. Statist. Soc. A 147, 278–292.

Gneiting, T., Raftery, A.E., Westveld, A.H., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon. Weather Rev. 133, 1098 – 1118.

Heffernan, J.E., Tawn, J.A., 2004. A conditional approach for multivariate extreme values. J. Roy. Statist. Soc. B 66, 497–546.

Hernandez, F., Smith, G., Baetens, K., Cossarini, G., Garcia-Hermosa, I., Drevillon, M., Maksymczuk, J., Melet, A., Regnier, C., von Schuckmann, K., 2018. New Frontiers in Operational Oceanography. Editors E. Chassignet, A. Pascual, J. Tintore and J. Verron. GODAE OceanView. chapter 29: Measuring performances, skill and accuracy in operational oceanography: new challenges and approaches. pp. 759–796.

Jonathan, P., Ewans, K.C., Randell, D., 2013. Joint modelling of environmental parameters for extreme sea states incorporating covariate effects. Coastal Eng. 79, 22–31.

Jonathan, P., Randell, D., Wu, Y., Ewans, K., 2014. Return level estimation from non-stationary spatial data exhibiting multi-dimensional covariate effects. Ocean Eng. 88, 520–532.

Messner, J.W., Pinson, P., Browell, J., Bjerregard, M.B., Schicker, I., 2020. Evaluation of wind power forecasts: an up-to-date view. Wind Energy 23, 1461–1481.

Murphy, C., Tawn, J.A., Varty, Z., 2024. Automated threshold selection and associated inference uncertainty for univariate extremes. arXiv preprint arxiv:2310.17999 .

Tendijck, S., Jonathan, P., Randell, D., Tawn, J.A., 2024. Temporal evolution of the extreme excursions of multivariate kth order Markov processes with application to oceanographic data. Environmetrics 35, e2834.

Tendijck, S., Ross, E., Randell, D., Jonathan, P., 2019. A non-stationary statistical model for the evolution of extreme storm events. Environmetrics 30, e2541.

Towe, R., Randell, D., Kensler, J., Feld, G., Jonathan, P., 2023a. Estimation of associated values from conditional extreme value models. Ocean Eng. 272, 113808.

Towe, R., Randell, D., Kensler, J., Feld, G., Jonathan, P., 2023b. Estimation of associated values from conditional extreme value models. Ocean Eng. 272, 113808.

Towe, R., Ross, E., Randell, D., Jonathan, P., 2024. covXtreme: MATLAB software for non-stationary penalised piecewise constant marginal and conditional extreme value models. Environ. Model. Softw. 177, 106035.

Towe, R., Zanini, E., Randell, D., Feld, G., Jonathan, P., 2021. Efficient estimation of distributional properties of extreme seas from a hierarchical description applied to calculation of un-manning and other weather-related operational windows. Ocean Eng. 238, 109642.

Winter, H.C., Tawn, J.A., 2016. Modelling heatwaves in central France: a case-study in extremal dependence. J. Roy. Statist. Soc. C 65, 345–365.