Review of manuscript egusphere-2026-3111 "Forecasting threshold exceedance of atmospheric variables at a specific location".

This manuscript compares two frameworks to obtain probabilistic forecasts for the occurrence of extreme events, defined as exceedances of a relevant climatological threshold. The first approach trains a model to directly forecast the conditional probability that an extreme event will occur (given some covariates), while the second approach learns the entire conditional distribution of the outcome, and then derives the threshold exceedance probability from the estimated distribution. In a toy example where the joint distribution of the covariates and outcomes is known, the authors derive analytical formulae for relevant verification metrics as the unconditional event probability tends to zero. These are compared for the two model classes, and it is found that the distributional approach performs better than directly modeling the threshold exceedances, particularly for very extreme events. The theoretical formulae are then verified using a simulation experiment, before being compared to results in an empirical application to neural network-based forecasting models for wind speed and precipitation at ground-stations in France.

Overall, the comparison is interesting, and the topic of the paper is relevant for the journal. The structure of the paper is sensible, with the theoretical results supported by empirical analyses in both simulated and then real-world applications. While several other papers have similarly studied whether distributional regression methods outperform binary classification methods, the main novelty of this manuscript is the derivation of the theoretical prediction errors within the toy model. However, this toy model is relatively simple, limiting the scope of the paper, and the theoretical results make some fairly strong assumptions that are not mentioned or discussed in the main text. At the very least, the presentation of the models, assumptions, and results should be improved so that it is clearer what the contribution of the paper is, and what its limitations are. These comments are explained in more detail below.

Major Comments

- It is remarked that "the primary purpose of this paper is to systematically compare direct binary classification versus full-distribution parametric modeling for predicting the likelihood of extreme events." Over the past couple of decades, several other papers have similarly compared forecasts from predictive distributions and probabilistic classifiers, particularly when interest is on extreme events, both in the atmospheric sciences and in many other domains. The paper would therefore benefit from a clearer explanation of where the novelty of the paper lies, and how it differs from existing studies on the same topic.

- In my opinion, the main novelty of the paper is the derivation of the theoretical squared errors of the forecasts from the two models as the probability tends to zero in equations 24 and 25, and the subsequent Brier skill scores and log score differences. However, these results are only valid for the toy example considered in Section 3, which is fairly simple. It could be made clearer in the text for what models these results are actually derived. For example, it could be stated more clearly that the model in class $\mathcal{M}_1$ corresponds to a logistic regression model that allows for non-linear relationships between the covariates and the logit transformed outcome, and that the model in class $\mathcal{M}_2$ corresponds to a Gaussian distribution with fixed variance and a mean that can be obtained using any point prediction method (with both models trained using maximum likelihood estimation). There are many other models in $\mathcal{M}_1$ and $\mathcal{M}_2$ for which the theoretical results do not apply, and this should be clarified. This would help to understand the scope of the results, and how they are expected to align with score differences in practical applications.

- Related to this, the toy model assumes that the outcome variable follows a Gaussian distribution with mean equal to a transformation of the covariates, and a constant variance. However, the applications in Section 4 do not align with this assumption: neither wind speed nor precipitation follows a Gaussian distribution, for example. It is therefore unclear what results we expect to obtain in these applications based on the theory in Section 3. As a result, this application is somewhat disconnected from the theory,

making it very similar to the many other comparisons of probabilistic classifiers and distributional regression methods.

- The derived theoretical results also require making some relatively strong assumptions, even within the toy model, and these are not mentioned in the main text. For example, before equation A9, it is assumed that there is no correlation between the estimated parameters, which is often unrealistic, and it is also assumed that V_1 and V_2 are independent of the covariates, which seems very limiting. While I understand that this yields the simplest expression of the squared errors, it should at least be discussed that these assumptions are being made, whether or not they are reasonable, and what would happen to the results if these assumption were not made.

  Moreover, it is mentioned (line 332) that "to achieve optimal alignment between theory and empirical results, we calibrated the constant terms in these analytical expressions," and "incorporating a quadratic correction term ... provides marginally better agreement." I do not find it obvious from the text what has been calibrated and corrected, and why this is necessary. In particular, this suggests that the theoretical results alone do not explain the score differences that are observed in practice, and this should be discussed in detail in the text.

Minor Comments

- There are several typos and grammatical errors in the text, which make the paper difficult to follow at times.
- Title: Similar to my first major comment above, the title is quite vague and could be changed to highlight the exact novelty of the paper. The term "at a specific location" also suggests that the paper is focused on forecasts of extreme events at a particular weather station, whereas this refers more to how extreme events are defined. Perhaps "site-specific threshold exceedances" would be more suitable?
- Line 50: "framing the task as a standard binary classification problem optimized via Binary Cross-Entropy (BCE)." This suggests that this "direct exceedance modeling" approach *must* be trained to optimize the BCE, but a probabilistic classifier can also be trained using the Brier score or any other proper scoring rule, for example.
- Line 59: It is mentioned that one "classification approach" is to take the numerical weather prediction (NWP) ensemble and estimate "the probability of exceedance from the fraction of physical members that surpass a target threshold." However, I would argue that this is a "full distribution" approach, since the NWP ensemble provides a sample from the (estimated) conditional distribution of the outcome variable given the information available at the time of forecasting.
- Lines 79–90: This paragraph lists several non- and semi-parametric probabilistic forecasting methods. However, these are not really relevant for the rest of the paper, where a parametric model is always assumed. I would recommend condensing this paragraph and including it in the previous paragraph on "techniques designed to predict the full probability distribution."
- Lines 117–120: Why are $V(t), T(t), R(t)$ introduced? These are not defined, nor are they used again in the paper. Moreover, why is $Y(t)$ written explicitly as a function of time, while all other variables denoted time as a subscript?
- Equation 1: It could be briefly mentioned here that the analysis also holds when interest is on threshold non-exceedances (i.e. extremely low outcomes).
- Line 126: The term "(with $p \ll 1$)" could be removed here. The following expression holds for all $p \in (0, 1)$, and it is clarified in the following sentence that interest is on $p \ll 1$.
- Equation 3: I would define p_t as the conditional probability of I_{t+h} occurring given X_t, rather than given $\mathcal{F}_t$. Of course, X_t can be interpreted as an encoding of $\mathcal{F}_t$ that is suitable for model training, but since the mathematical analysis treats p_t as the conditional probability of I_{t+h} occurring given X_t, it may be better to explicitly define it like this here.

- Line 145: It is not really clear what it means for M_1 to be an element in $\mathcal{M}_1$, since $\mathcal{M}_1$ has not formally been defined.
- Equation 6:
  - A "+" is missing between the two terms in this expression.
  - It may be useful to point out that the BCE is just the average log score for a probabilistic classifier.
  - What is the motivation for including a subscript t everywhere, when this is itself indexed by i? For example, what is the difference between $\hat{p}_t^{(1)}$ and $\hat{p}_{t_i}^{(1)}$? On line 254, the subscript $t = 1, \ldots, N$ is used without any sub-indexing, and again in the first equation of Appendix A. A similar argument can be made for the forecast horizon h in Section 4.2.
  - It could perhaps be mentioned here that $X_1, \ldots, X_N$ (or $X_{t_1}, \ldots, X_{t_N}$) represent a (random) training data set on which the model parameters are learned. This is important to clarify when interpreting the randomness in $\hat{\boldsymbol{\theta}}$ in the theoretical results.
- Line 151: "The BCE loss is particularly suitable for this task as it directly optimizes for probability calibration." I would refrain from using the term "probabilistic calibration" here, since there is no guarantee that a model trained using BCE issues calibrated or reliable predictions (at least when the model class is not well-specified). Hence, it arguably does not "directly optimize for probability calibration".
- Line 156: "Following parametric deep learning frameworks..." This modeling approach is also used in simple statistical models, it does not need to correspond to a deep learning framework.
- Line 160: If $f(y; \Pi_t)$ is the conditional density of $Y(t+h)$ given $\mathcal{F}_t$, as suggested in the text, then it is assumed that the model class is well-specified, and that Π_t is the "true" parameter vector underlying the data-generating process. This could be clarified in the text to avoid confusion. If this is not assumed, then $f(\cdot; \hat{\Pi}_{t_k})$ in Equation 8 could perhaps be replaced by $\hat{f}(\cdot; \hat{\Pi}_{t_k})$, which would clarify that the form of the density is also being estimated.
- Line 161: Perhaps mention here that the parameters do not need to be learned by "minimizing the negative log-likelihood." As mentioned in the introduction (line 72), other proper scoring rules could also be used.
- Equation 8: An index k is used here, whereas the index i is used in Equation 6. Is there a reason for this? If not, I would use consistent notation throughout.
- Line 171: It may be useful to explain what is meant exactly by the term "heavy class imbalance". This will likely be obvious to some readers, but not to others.
- Line 173: I assume $I_t(Y_0)$ should be $I_{t+h}(Y_0)$, or simply I_{t+h}? The interpretation does not change, but this would keep the notation consistent.
- Line 173: The terms TP, TN, FP, and FN do not need to be defined mathematically, but I would at least mention what they stand for. Again, I imagine this will be obvious to many readers, but not to others.
- Equation 14: Why introduce the Brier skill score but not the Log skill score?
- Line 198: Perhaps explain here what a proper scoring rule is, or why it is beneficial. Moreover, while the reference Gneiting et al. (2006) does mention proper scoring rules, it is not the focus of the paper; perhaps the authors meant to cite Gneiting and Raftery (2007) instead?
- Line 199: It could perhaps be mentioned how the AUC differs from scoring rules. This would help to motivate including it in addition to the Brier score and Log score.
- Line 225: The term "reciprocal function" is not standard terminology for the inverse function. I would avoid this term to avoid confusion.

- Equation 18: The z in the expectation should be capitalized.
- Line 217: In $(X_t)_{t\in\mathcal{T}}$, what is $\mathcal{T}$ and why is it used here but nowhere else in the paper?
- Line 254: Perhaps avoid using square brackets for $[X_t, Y(t+h)]$ since this makes the pairs look like an interval, which they are of course not.
- Line 260: $p_t^{(k)}$ should be $\hat{p}_t^{(k)}$.
- Equations 24 and 25:
  - I would explain what the notation $\underset{p\to 0}{\sim}$ means. I assume this will be unfamiliar to many readers.
  - It is quite difficult to distinguish between p and ρ in these equations. Consider using a different symbol to represent the noise-to-signal ratio.
  - It should be noted that the expectations in these equations are over both the random training sample and the random covariate value. This notation varies throughout the paper, with some expectations clarifying which variables they are over, and other expectations not.
  - What are K_1 and K_2? It should at least be mentioned that these are functions of the noise-to-signal ratio that do not depend on p, but do depend on the chosen model used to obtain the distributional parameters.
  - After the equation, it is mentioned that "this result first indicates that, at fixed p (small enough), as the noise-to-signal ratio increases, estimation error decreases." However, this is not possible for readers to deduce without knowing how the functions K_1 and K_2 behave, since these are also functions of the noise-to-signal ratio.
- Line 268: "assuming that both M_1 and M_2 estimation methods are efficient." What is meant by the term "efficient"? The results in the appendix suggest that it is assumed the estimation methods are consistent. Is this what is meant?
- Line 270: "This result originates from the fact that the effective amount of "information" used to calibrate M_1 parameters is pN." Does this result still hold if we consider $p \to 0$ but $pN \to \infty$? The text suggests that the result is inherent in the finite sample size on which the methods are trained. It would be useful to clarify this. On line 284, the results are considered for $N \to \infty$.
- Lines 272–300: What is the motivation for studying the Brier skill score, but then the difference in log scores, and the Peirce skill score ratio? These are three different ways to compare the two forecasts. A log skill score can be defined analogously to the Brier skill score, so why is this not considered? It makes these results seem somewhat ad-hoc and like a random collection of results, as if the authors have chosen the comparison for each score so that the results fit the storyline best. I assume this is not the case, but choosing one general framework to compare the two forecasts (e.g. by looking at ratios of skill scores) would be easier to motivate, and would give the results more structure.
- Line 283: "For PSS, we establish ... explicit expressions in terms of p_t-averaged values for methods M_1 and M_2. Such integrals that can be evaluated numerically." I do not understand this sentence, consider rephrasing.
- Line 284: "In the regime $N \to \infty$, we obtain the the following asymptotic behavior for $p \to 0$." Does this not also depend on the limiting behaviour of pN?
- Equations 28 and 29: I assume C_1 and C_2 are positive constants that do not depend on p or ρ? This should be clarified.
- Line 298: The term K_p in this equation looks very similar to the term K_ρ in equations 27–29, which itself employs different notation to $K_1(\rho)$ and $K_2(\rho)$ in equations 24 and 25. The presentation of these results could be adapted to make the differences between these terms clearer.
- Line 312: What is meant exactly by the term "logits loss"?

- Line 316: Why is $d = 12$ chosen? Does this choice have any impact on the results?
- Figure 1: Since we are interested in comparing the error of the two models, I would plot the two panels of Figure 1 with a common y-axis. From this, it should be clear that M_2 outperforms M_1.
- Line 326: "We have verified that squared bias terms are negligible in the regime we consider, thereby validating the variance-dominated error assumption in Appendix A for the considered range of exceedance probabilities." What is meant by this exactly?
- Line 331: "The dashed curves in panel (a) and solid curves in panel (b) represent our analytical predictions derived from Eqs (A19) and (A9) respectively." Since these equations in the appendix have not yet been mentioned in the text, it would be useful to clarify how they correspond to the results already presented, such as equations 24 and 25. It is difficult to interpret the meaning of these curves without more context.
- Line 333: "It is noteworthy that, for the ε_1 case when $\rho_2 = 1$, incorporating a quadratic correction term $V_1'(\mu) = V_1' + V_1''\mu^2$ in Eq. (A19)." Firstly, these terms have not yet been introduced or mentioned in the text, making it difficult to interpret what this sentence implies without readers looking through all of the technical details in the appendix. Secondly, this quadratic correction suggests that the theoretical results need to be corrected, or cannot be calculated exactly; this should be explained and justified in the main text.
- Line 335: "the results demonstrate excellent concordance between the estimated data and our analytical expressions, thereby providing empirical validation for theoretical hypotheses of Appendix A." It is not really discussed what these hypotheses in Appendix A are, which makes it difficult to interpret this result.
- Line 362: The results are presented for neural network prediction models. ML models generally require much more training data than more parsimonious statistical models, which means they are particularly effected by the reduced effective sample size for extreme events. Such models generally do not fit into the class of models considered in the theoretical analysis (at least not for $\mathcal{M}_1$, which apply a sigmoid function to some mean prediction). How are the results expected to change in this case?
- Line 379: "During this first stage the integrity of data is preserved and no quality check is applied." The integrity of the data is preserved and no quality check is applied? What is meant by this?
- Line 381: What is meant by the term "to ensure statistical significance"? The "keys" have not yet been introduced, making this sentence difficult to understand. For example, what are the thresholds of 2000 and 500 applied to?
- Section 4.1: I found this section quite difficult to follow. It could benefit from a clearer presentation of the models, observations, and what is to be done in the analysis. For example, it could already be clarified here that the NWP models are used as inputs to two ML prediction models belonging to $\mathcal{M}_1$ and $\mathcal{M}_2$.
- Table 1: It could perhaps be clarified here (e.g. in the caption) that the spatial grid refers to the grid around a station of interest. Moreover, since the forecasts are considered at ground-station observations, what is a "local subgrid of the 2-D and 3-D NWP fields centered on the station"? In particular, how is the third dimension of the ARPEGE fields used?
- Line 406:
  - The conditional probability is denoted here by $P(Y \leq y \mid Y > 0)$. Previously, the probability was denoted by Prob rather than P. Similarly, Y is typically denoted as a function of time.
  - I also assume that $F^S_{C,+}(Q_p) = p_{+,R}$ should be $F^S_{C,+}(Q_p) = 1 - p_{+,R}$? Moreover, what does the subscript R represent? The text suggests that the quantiles are calculated separately for each station, so should this be a subscript S instead?

- Line 409: The probability $\bar{p}_{wet}$ does not have a bar in Appendix E (e.g. equation E2).
- Line 426: "The parameter ν controls the mean level." This parameter is itself determined by two other parameters. Why are these not mentioned here?
- Line 447: The "key" encodes the "station S, day d, and time t." What is the difference between the day and the time? Since hourly observations are considered, I assume the time corresponds to the hour of the day?
- Line 451: I assume $\boldsymbol{D}_S$ should be $\boldsymbol{D}_{\mathrm{k}}$?
- Line 453: What is n_S? Does the number of input variables at each target site differ? Similarly, what does n_{AR} represent on line 457?
- Line 466: The notation $\boldsymbol{AR}_{d,t}^S$ and $\boldsymbol{AP}_{d,t}^S$ differs from that employed previously. I assume these should be $\boldsymbol{AR}_{\mathrm{k}}$ and $\boldsymbol{AP}_{\mathrm{k}}$?
- Line 481: "all forecasting tasks and labels corresponding to the same calendar day, which exhibit the strongest autocorrelation, were strictly assigned to the same subset." What is meant by this? This suggests that the correlation between the same day in consecutive years is stronger than the correlation between consecutive days in the same year, which seems unlikely.
- Line 484: "the complete pool of potential calendar days was randomly partitioned into training (85%), validation (10%), and test (5%) subsets." With 5% of three years of data, this suggests the (effective) test data set is quite small, even when augmented with the validation data. How much uncertainty is there in the estimated score differences?
- Line 486: "As reported in Table 1, ..." Information about the total number of keys is not reported in Table 1.
- Line 525: "the term $\frac{Rel}{U}$, even if increasing rapidly with $p \to 0$, remains relatively well controlled." What is meant by the term "well controlled"?
- Table 3: Why are the LS and AUC presented in a table while all other metrics are presented in a figure? The argument that "these metrics are not suitable for objective comparison across different base rates" also holds for the Brier score, but the skill score (which can also be applied to the LS) circumvents this.
- Figure 7: Personally, I would not include the results for the CSI and HSS in the main text. There are no theoretical results to support these empirical results, unlike for the other scores, and the analogous conclusions are drawn, meaning this analysis does not add anything to the paper. I would move these to an appendix instead, but I leave this to the discretion of the authors.
- Line 570: $\mathcal{F}$ is used here to denote a parametric family, whereas it is previously used to denote the information available to the forecaster.
- Line 574: While I agree that proper scoring rules should be used to compare competing probabilistic forecasts, I do not agree with the statement that "without access to this ground truth, the only operational proxy for reality is the aggregated evaluation of proper scoring rules ... averaged over a heterogeneous test set." In what sense is an average score an "operational proxy for reality"? Moreover, once we acknowledge that our model class is (almost) always misspecified, different loss functions (or scoring rules) will generally yield different forecasts, meaning there is not a direct correspondence between the model space and reality.
- Line 599: These results hold for all metrics considered, not just when predictive skill is "measured by the PSS".

- Line 613: "The success of distributional models intrinsically does not relies so much on the suitability of the chosen parametric family." This only holds within reason. The distributions considered for modeling rainfall and wind speed were all relatively similar. If we were to use a completely different, inappropriate distribution, then I assume approach $\mathcal{M}_2$ would generally perform worse than $\mathcal{M}_1$. I do not agree that this is improvement is "intrinsic" in all distributional models.
- Line 620: What is meant by a "continuous spatial domain"? All spatial domains considered in practice are in some sense discretized.
- Appendices: The notation used throughout the appendices (particularly Appendix A) is quite inconsistent, making the theoretical results difficult to follow. For example:
  - Line 643: The main text introduces $\hat{p}_t^{(1)}$ and $\hat{p}_t^{(2)}$ as the estimated exceedance probabilities, but $\hat{\mu}$ has not yet been defined. On the following line, $\hat{\mu}(Z;\theta)$ is used, whereas later $\hat{\mu}(X_t)$ is used. What is Z here?
  - Line 644: It should perhaps be mentioned that the Gaussian log-likelihood is only equivalent to the MSE when the variance is assumed to be known. In many applications (including those in Section 4 of the manuscript), the variance parameter is estimated as well, in which case this equivalence does not hold.
  - Line 648: $\nabla^2_{\boldsymbol{\theta}}\ell(Y,X;\boldsymbol{\theta}_0)$ should be $\nabla^2_{\boldsymbol{\theta}}\ell(Y,X;\boldsymbol{\theta})\mid_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$. Similarly for the definition of $G(X)$ on line 651.
  - $\hat{M}_1(X_t;\boldsymbol{\theta}), \hat{M}_2(X_t;\boldsymbol{\theta}), \hat{z}(X_t;\boldsymbol{\theta})$ are sometimes written with a semi-colon and sometimes with a comma. The dependence on $\boldsymbol{\theta}$ is also sometimes suppressed, but not always.
  - Equation A6: What exactly does $\approx$ mean here? The symbol $\backsimeq$ is also used later in the manuscript (e.g. before equation A20 and in equation E1). Is there a difference between them? This should be an asymptotic variance, which is not apparent from the current notation. Moreover, this variance is conditional on X_t, which is also not clear.
  - Line 665: μ_t is itself a random variable, so it is not clear what the conditional expectation $\mathbb{E}(V_2'(X_t) \mid \mu(X_t) = \mu_t)$ actually represents. Is this meant as the conditional expectation given the σ-algebra generated by μ_t?
  - Equation A8: In this integral, $V_2'(u)$ is the conditional expectation given $\mu(X_t) = \mu_t$ defined above, rather than the original function V_2' defined in equation A5, right? The notation does not allow these to be distinguished.
  - Equation A8: The exponential in this integral is derived from substituting $Q_p = -\sqrt{s^2+\sigma^2}\Phi^{-1}(p)$ and $\rho^2 = \sigma^2/s^2$, yielding $Q_p = -(\sigma/\rho)\sqrt{1+\rho^2}$. The factor σ/ρ then seems to have been taken out of the term $(Q_p - u)^2$ without also rescaling u? Why can this be done? I assume I have missed something?
  - Line 672: What is meant by $\mathbb{E}(V_2'(\mu)) \approx V_2$? The left-hand-side is a single value, whereas the right-hand-side is a function.
  - Equation A9: V_2 is defined as a function but is here treated as a constant?
  - Equation A16: What is $\hat{p}^{(3)}$?
  - Appendix C: Many of the integrals in this section are missing dx.
  - Line 730: The notation $\mathbb{E}_{I_{t+h}|p_t}$ has not been defined, and is different from the notation for conditional expectations and probabilities elsewhere in the paper (e.g. line 665).
  - Line 746: "Assuming consistency of the estimators, ie., that $\hat{p}_t^{(k)} = p_t + \delta_t$ with $\delta_t \ll 1$." This is not the standard definition of the consistency of an estimator. On line 841, δ is also instead used to denote a Dirac measure.
  - Lines 765–770: $P(X_t)$ is defined as the conditional expectation of I_{t+h} given X_t. I assume $P(x)$ is then the conditional expectation of I_{t+h} given $X_t = x$? This should be clarified.
  - Line 806: I assume Z should be X? Moreover, $X_t \in \mathbb{R}^{N\times d}$ suggests that X_t is a matrix, but X_t is actually a d-dimensional vector.

– Equation E1: How does n in this equation differ from N used elsewhere? What is H_{n-1}? What are the y_i's? It should be clarified that this is a sample estimate of the mean, rather than an analytical representation of it.

– Appendix E2: The parameter vector $\boldsymbol{\theta}$ is not bold in this section, in contrast to in all other sections.

References

Gneiting, T., Larson, K., Westrick, K., Genton, M. G., and Aldrich, E. (2006). Calibrated probabilistic forecasting at the stateline wind energy center: The regime-switching space–time method. *Journal of the American Statistical Association*, 101:968–979.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378.