

Complete Response to the Referee 1

Manuscript egusphere-2026-3111

Roberta Baggio and Jean-François Muzy

We sincerely thank the referee for the careful reading of the manuscript and for the detailed and constructive comments. We believe that addressing these comments has substantially improved the manuscript, in particular by clarifying the scope and contribution of the study, its relation to the existing literature and the assumptions and limitations of the theoretical analysis.

The revision has involved substantial changes throughout the manuscript. Before addressing the comments individually, we summarize the main revisions. First, the Introduction has been reorganized to better contextualize the comparison between direct classification and distributional modeling, to acknowledge previous empirical comparisons more explicitly and to identify more precisely the contribution of the present study. Second, Section 3 has been revised to state more clearly the role of the Gaussian location model, the assumptions and approximations required for the analytical results and the limits of their applicability. Finally, the Discussion and Conclusions have been substantially revised to emphasize the two complementary aims of the study: (i) to characterize, through the analytical estimation of the forecast errors within an idealized framework, the loss of statistical information induced by thresholding a continuous observation and (ii) elucidate simple statistical mechanisms contributing to the predictability of rare threshold exceedance in real weather data by considering analogies and departures from this theoretical model. The detailed changes made in response to each comment are described in the following.

Major comments

Referee comment. 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.

Response. We agree that both paradigms are established and that empirical comparisons exist. Our original wording was too broad. We now cite, among others, ?, who compare point-wise probability forecasts with approaches based on full predictive distributions for precipitation, ?, who explicitly consider the performance of both paradigms for the prediction of extreme events and ?, who provide a recent comparison of traditional and machine-learning-based probabilistic post-processing methods. It is now more explicit that we do not introduce the paradigms or provide the first empirical comparison. Our contribution is narrower: we place the two strategies in a common controlled framework, derive how their probability-estimation errors behave as the event base rate decreases and propagate these errors to forecast scores. The mechanism isolated by the toy model is information loss: the direct classifier is trained after Y has been replaced by $\mathbf{1}\{Y > Q_p\}$, whereas the distributional estimator gathers information from all observations. The applications then examine whether the qualitative mechanism remains visible in real meteorological data with substantially more complex statistical properties. The observed agreement is consistent with the hypothesis that predictable shifts in the conditional distribution contribute to the forecastability of the extreme exceedances considered here.

Referee comment. 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.

Response. We thank the Reviewer for this observation and agree that the distinction between the general model classes and the specific assumptions of the theoretical analysis was not sufficiently clear. We have revised the manuscript accordingly. In our general formulation, $\mathcal{M}_1$ denotes models that directly estimate the probability of a thresholded binary response, whereas $\mathcal{M}_2$ denotes models that retain the continuous predictand and estimate its conditional distribution. Neither class is intrinsically restricted to the specific functional forms considered in Section 3. The theoretical analysis compares specific representatives of these two classes for estimating threshold-exceedance probabilities generated by a Gaussian random-location model. The particular model $M_2 \in \mathcal{M}_2$ used in the analysis is the corresponding correctly specified Gaussian conditional model, whose mean can be represented by a flexible nonlinear function of the covariates. The model $M_1 \in \mathcal{M}_1$ directly estimates the Bernoulli probability obtained by thresholding the same Gaussian response. Its logit is likewise allowed to depend nonlinearly on the covariates. Both models are correctly specified for their respective responses, target the same conditional exceedance probability and are estimated by maximum likelihood under the regularity assumptions stated in Appendix A. This construction allows us to isolate the information loss caused by replacing the continuous observations with threshold-specific binary labels. The explicit prediction-error expressions and score asymptotics derived in Section 3 consequently apply to this Gaussian random-location model and to the associated regular estimation framework. They are not claimed to hold for arbitrary members of $\mathcal{M}_1$ and $\mathcal{M}_2$. The meteorological applications subsequently examine whether the qualitative mechanism isolated in this idealized setting remains visible for non-Gaussian, heteroskedastic, and potentially misspecified weather variables. These limitations and the precise scope of the theoretical results are now stated explicitly in the revised manuscript.

Referee comment. 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.

Response. We agree that the Gaussian location model in Section 3 does not provide a full quantitative description of the wind-speed and precipitation applications. Wind speed and precipitation are non-Gaussian, their conditional dispersion and shape may vary with the predictors, and the parametric models used in Section 4 may be misspecified. We have revised the Introduction, the end of Section 3.1, the opening of Section 4, the Discussion, and the Conclusion to state this distinction explicitly. The Gaussian model is used as an analytically tractable proof of concept in which the information loss caused by thresholding a continuous response can be isolated, quantified, and checked numerically. The derived powers of the exceedance probability p are specific to this Gaussian location model with known constant conditional variance. We no

longer present the empirical applications as direct validations of these quantitative asymptotic formulas. The connection with Section 4 is instead formulated at the level of the underlying mechanism. When extreme exceedances are substantially associated with predictable changes in the bulk of a conditional distribution, a distributional model can learn these changes from continuous observations over the full response range. A direct classifier discards the magnitude information and must learn separately from the increasingly sparse binary labels associated with each threshold. The qualitative expectation tested in Section 4 is therefore that the relative advantage of distributional modeling should increase as the threshold becomes more extreme. For wind speed, both the increasing discrimination of M_2 and its growing relative advantage are recovered. For rainfall, PSS_2 remains approximately constant rather than increasing, but the relative advantage of M_2 over M_1 still widens toward rarer thresholds because the classifier deteriorates more strongly. We have also revised the discussion of the parametric family. The similar performances obtained with the alternative precipitation families are now interpreted only as evidence that, for the datasets considered, families, and threshold ranges considered, capturing predictable bulk changes appears more important than the precise choice among these reasonably suitable parametric families. We explicitly state that this does not imply that distributional shape or tail specification is unimportant in general. Severe misspecification, covariate-dependent tail changes, heteroskedasticity, or non-negligible estimation bias may alter the conclusions. Finally, we identify extensions to non-Gaussian location-scale models, covariate-dependent dispersion, mixed discrete-continuous distributions, and misspecified models as important directions for future theoretical work.

Referee comment. 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.

Response. We thank the Reviewer for pointing out that the assumptions underlying the analytical results and the interpretation of Fig. 1 were not sufficiently explicit. We have revised both the main text and substantially rewritten Appendix A accordingly. First, the general delta-method expressions do not assume uncorrelated parameter estimates: they involve the full inverse Fisher information matrix, including its off-diagonal terms. For M_1 , the closed-form approximation additionally relies on a factorization that neglects the matrix-valued covariance between the conditional Bernoulli variance $p_t(1 - p_t)$ and the local parameter-sensitivity matrix. This factorization does not diagonalize the Fisher matrix. Without it, the general error expression remains valid but requires numerical evaluation of the full Fisher matrix and the corresponding quadratic forms. Second, the closed-form results approximate the conditional local variance functions by positive constants,

$$V_{1,p}(X_t) \simeq V_1, \quad \bar{V}_2(u) \simeq V_2.$$

These are modeling approximations, not assumptions of statistical independence. More general covariate-dependent functions could be retained, but would introduce additional model-specific coefficients. Numerical diagnostics based on the conditional prediction variances show only weak variation of the reconstructed local variance functions over the ranges contributing most to the errors, supporting the constant approximation in the present experiments. For M_1 , the revised

finite- p expression now retains the exact scalar Bernoulli-variance contribution within the factorized approximation, $C_p = \mathbb{E}[p_t(1 - p_t)] = p - \Phi_2(\Phi^{-1}(p); r)$, with $r = (1 + \eta^2)^{-1}$, instead of replacing it directly by p . This improves the finite- p description without changing the leading rare-event scaling. We also clarified the comparison with the numerical experiments. The previous references to “calibrated constants” and a “quadratic correction” have been removed. In the revised figures, one positive factor, V_1 or V_2 , is fitted for each method and value of η^2 . These factors depend on the network architecture, parameter covariance, output sensitivity, and optimization procedure, and cannot be determined from the scalar Gaussian data-generating model alone. Accordingly, the experiments assess the predicted functional dependence on p , rather than a parameter-free prediction of the absolute error amplitudes. Finally, we now state explicitly that the results assume a correctly specified Gaussian location model with known conditional variance, regular asymptotic normality, negligible squared bias, the stated Fisher factorization for M_1 and approximately constant local variance functions. The large-sample results are first derived for fixed $p > 0$ as $N \rightarrow \infty$, and their rare-event behavior is then examined as $p \rightarrow 0$. If $p = p_N \rightarrow 0$ jointly with N , regular asymptotic normality for M_1 additionally requires $Np_N \rightarrow \infty$.

Minor and presentation comments

1. **Referee comment.** There are several typos and grammatical errors in the text, which make the paper difficult to follow at times.

Response. The manuscript has been extensively revised for language, presentation, and notation. We have provided a full latex difference document.

2. **Referee comment.** 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?

Response. We agree that the previous title was too vague and that the phrase “at a specific location” could be misinterpreted as referring to a particular weather station. We have therefore removed this expression and revised the title to emphasize the two complementary contributions of the paper: the analytical study of rare-event forecasting and its application to weather variables. The revised title is: “Forecasting rare exceedance events: analytical insights and weather applications.”

3. **Referee comment.** 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.

Response. We agree and we have removed the statement suggesting that direct exceedance models must be trained using BCE. The revised text defines the class more generally as models that directly estimate the exceedance probability without restricting the training criterion to BCE. BCE loss is subsequently presented only as the training loss used in our implementation.

4. **Referee comment.** 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.

Response. We agree that the paragraph was not sufficiently clear and was mixing different problems, the binary outcome and the type of data and techniques used to resolve this task. In the new version of the introduction, these two distinct aspects are more clearly separated and more relevant citations are given for each specific aspect.

5. **Referee comment.** 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.”

Response. We agree on this point and this paragraph has been shortened significantly. Examples of non-parametric and semi-parametric methods are cited for completeness, but it is said more explicitly that the parametric approach is the one really relevant for the study.

6. **Referee comment.** 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?

Response. The undefined and unused variables $V(t)$, $T(t)$, and $R(t)$ have been removed. We have also harmonized the notation for time-dependent variables throughout the manuscript and now use subscripts consistently, for example Y_t rather than $Y(t)$.

7. **Referee comment.** 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).

Response. Done.

8. **Referee comment.** 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 $p \ll 1$.

Response. Done.

9. **Referee comment.** 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.

Response. We agree and have followed the reviewer’s suggestion. In the revised manuscript, the conditional exceedance probability is defined directly as $p_t = \Pr(I_{t+h} = 1 \mid X_t)$, where X_t denotes the observable covariate vector used by the forecasting models. This definition is now used consistently throughout the theoretical analysis and we no longer refer to $\mathcal{F}_t$.

10. **Referee comment.** 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.

Response. We have rephrased the sentences in Subsections 2.2.1 and 2.2.2 to make more explicit that $\mathcal{M}_i$ are sets of models while M_i denotes a specific realization within the set.

Referee comment. 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.
- 11. • 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 $t = 1, \dots, N$ 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 $X_1, \dots, X_N$ (or $X_{t_1}, \dots, X_{t_N}$) represent a (random) training data set on which the model parameters are learned. This is important to clarify when interpreting the randomness in $\hat{\theta}$ in the theoretical results.

Response. We thank the reviewer for these detailed comments. The missing “+” sign in the BCE expression has been restored. We now state explicitly that the BCE is the average logarithmic score for a binary probabilistic forecast. We have also harmonized the sample indexing: t_i identifies the time associated with training sample i , so that $\hat{p}_{t_i}^{(1)}$ denotes the prediction for that sample. Finally, We now identify the random training samples explicitly, $\{(X_{t_i}, I_{t_i+h})\}_{i=1}^N$ for M_1 and $\{(X_{t_i}, Y_{t_i+h})\}_{i=1}^N$ for M_2 , to estimate $\hat{\theta}$, whose randomness in the asymptotic analysis therefore arises from repeated training samples.

- 12. **Referee comment.** 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”.

Response. We agree and have removed the statement that BCE directly optimizes probability calibration. BCE is now presented as the strictly proper binary scoring rule used in our implementation, although other proper scoring rules could also be considered. Strict propriety is a population-level property and does not guarantee finite-sample calibration under model misspecification. Calibration must therefore be assessed empirically.

- 13. **Referee comment.** 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.

Response. We agree and this sentence has been eliminated.

- 14. **Referee comment.** 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.

Response. We agree that the previous definition implicitly assumed correct specification by identifying $f(y; \Pi_t)$ with the true conditional density. We have therefore removed the definition $f(y; \Pi_t) := \frac{d}{dy} \text{Prob}(Y(t+h) \leq y \mid \mathcal{F}_t)$. In the revised manuscript, $f(y; \Pi)$ denotes

the density of a chosen parametric family used to approximate the conditional distribution of Y_{t+h} given X_t , without assuming at this stage that the family contains the true conditional law. Only the time-dependent parameters $\hat{\Pi}_t = M_2(X_t; \hat{\theta})$ are estimated, so we retain the notation $f(y; \hat{\Pi}_t)$ rather than $\hat{f}(y; \hat{\Pi}_t)$. Correct specification is invoked only in the Gaussian theoretical analysis of Sect. 3 and Appendix A

15. **Referee comment.** 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.

Response. Done, the sentence makes now clear that minimization of negative log-likelihood is just one among other possible approaches.

16. **Referee comment.** 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.

Response. Corrected, there was no reason to distinguish between k and i .

17. **Referee comment.** 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.

Response. The meaning of heavy-class imbalance in the considered setting has been clarified, we now state explicitly that we mean very low event base rates.

18. **Referee comment.** 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.

Response. Corrected.

19. **Referee comment.** 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.

Response. Done: This section has now been expanded to make more explicit the meaning of the component of the confusion matrix.

20. **Referee comment.** Equation 14: Why introduce the Brier skill score but not the Log skill score?

Response. We agree that a climatology-referenced skill score can also be defined for the logarithmic score. We retain the BSS because it is widely used in meteorological verification and provides a clear measure of probabilistic skill relative to climatology. The logarithmic score is used as a complementary strictly proper score, and the additive difference $LS_1 - LS_2$ is particularly natural because it equals the difference between the corresponding expected Kullback–Leibler divergences. We have added this comment to the revised manuscript.

21. **Referee comment.** 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?

Response. Done: We corrected the reference and an explanatory sentence precisising the definition of a proper scoring rule has been added.

22. **Referee comment.** 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.

Response. Done: We now explicitly state that AUC is an additional measure of discrimination.

23. **Referee comment.** Line 225: The term “reciprocal function” is not standard terminology for the inverse function. I would avoid this term to avoid confusion.

Response. Corrected.

24. **Referee comment.** Equation 18: The z in the expectation should be capitalized.

Response. Indeed, this has been corrected.

25. **Referee comment.** 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?

Response. Corrected.

26. **Referee comment.** 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.

Response. We agree and now use parentheses to denote the training pairs, namely $\{(X_{t_i}, Y_{t_i+h})\}_{i=1}^N$, rather than square brackets.

27. **Referee comment.** Line 260: $p_t^{(k)}$ should be $\hat{p}_t^{(k)}$.

Response. Solved, this sentence has been changed in the revised version.

28. **Referee comment.** Equations 24 and 25:
- I would explain what the notation $\underset{p \rightarrow 0}{\sim}$ 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.

Response. We thank the reviewer for these comments. Equations 24 and 25 have been slightly revised and some clarifications have been added. (i) We now explicitly define the asymptotic notation immediately after these equations and at the beginning of Appendix A: $f(p) \sim g(p)$ as $p \rightarrow 0$ means that $f(p)/g(p) \rightarrow 1$ as $p \rightarrow 0$. (ii) Following the reviewer’s suggestion, we have replaced ρ by η throughout the manuscript to denote the noise-to-signal

ratio. This avoids the visual ambiguity between ρ and the exceedance probability p . (iii) We now state explicitly that the expectations defining $\mathcal{E}_1$ and $\mathcal{E}_2$ are taken jointly over repeated random training samples and over the covariate distribution. Under the negligible-bias assumption, these quantities correspond to the integrated asymptotic variances of the probability estimators. We have revised the notation in the relevant derivations to identify the variables being averaged whenever this is needed for clarity. (iv) The quantities $K_1(\eta)$ and $K_2(\eta)$ are now defined in Appendix A and described in the main text as positive prefactors that are independent of p for fixed η . They contain the model-dependent local variance factors V_1 and V_2 and therefore depend on the model architecture. Their values cannot be inferred from the scalar Gaussian generative model alone. In Fig. 1, one positive model-dependent factor is consequently fitted separately for each method. (v) We agree that the decrease of the errors with increasing η^2 cannot be inferred from the powers of p alone, because the prefactors $K_1(\eta)$ and $K_2(\eta)$ also depend on η . We now state that, from their explicit finite p expressions in Appendix and finite p , it can be deduced that PSS is decreasing for η^2 large enough. This is confirmed by numerical results and interpreted through

$$\text{Var}(p_t) = \Phi_2(z_p; r) - p^2, \quad r = (1 + \eta^2)^{-1}.$$

As η^2 increases, r tends to zero and $\text{Var}(p_t)$ vanishes, so that p_t approaches the constant marginal probability p . This explanation is now clearly distinguished from the rare-event scaling derived in Eqs. 24 and 25.

29. **Referee comment.** 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?

Response. We agree that the term “efficient” was ambiguous. The derivation assumes that the estimators are consistent and asymptotically normal and, under correct specification, have asymptotic covariance equal to the inverse Fisher information matrix. We have replaced the term “efficient” by this explicit statement in Sect. 3 and clarified the corresponding regularity assumptions in Appendix A.

30. **Referee comment.** 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 $p \rightarrow 0$ but $pN \rightarrow \infty$? 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 $N \rightarrow \infty$.

Response. We agree with the reviewer that this needs to be clarified. The quoted sentence has been replaced in the revised text by a more precise Fisher-information statement. Within the hypotheses of Appendix A, the total scalar information available to M_1 is proportional to

$$NC_p = N [p - \Phi_2(z_p; r)],$$

whereas the information used to estimate the conditional-distribution parameter in M_2 is proportional to N . Our results remain valid in a joint limit where $N \rightarrow \infty$, $p = p_N \rightarrow 0$, and $Np_N \rightarrow \infty$. The latter condition ensures that the expected number of exceedances diverges and that the regular asymptotic normality assumed for M_1 remains applicable. In this regime, the estimation errors of both methods vanish but the information available to M_1 grows more slowly than that available to M_2 . Therefore, the relative advantage of M_2 is not a consequence of keeping the training sample size fixed. We have clarified in Sect. 3 and Appendix A that the large-sample expressions are first derived for fixed $p > 0$ as $N \rightarrow \infty$, after which their rare-event behavior is examined as $p \rightarrow 0$. If p decreases jointly with N , the regular asymptotic framework for M_1 requires $Np_N \rightarrow \infty$. Regimes with a bounded or vanishing expected number of exceedances are outside the scope of the analysis.

31. **Referee comment.** 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.

Response. We thank the Reviewer for highlighting a weakness in the presentation: the rationale behind the different quantities used to compare the two models was not sufficiently explained. First, BSS and PSS are particularly convenient for comparisons across increasingly rare events because they remove, or avoid, much of the explicit dependence on the event base rate. In contrast, the raw Brier score is intrinsically affected by event prevalence. Concerning the fact that we only show PSS ratio, In our preliminary analysis we also considered the relative quantity $BSS_{21} = 1 - \frac{BS_2}{BS_1}$ which directly measures the relative improvement of M_2 over M_1 in terms of Brier score. For completeness, the corresponding results are shown below in Figure 1. However, we did not retain this additional quantity in the revised manuscript because it provides essentially the same qualitative information as the relative PSS comparison: BSS_{21} remains positive over the considered range of p , consistently indicating better performance of M_2 . Concerning the logarithmic comparison through $LS_1 - LS_2$, we have chosen this formulation as it is particularly natural from an information-theoretic perspective. The expected log score is the cross-entropy between the true and forecast distributions, which decomposes into the entropy of the true distribution plus a Kullback-Leibler divergence. Consequently, $LS_1 - LS_2$ directly equals the difference in the KL divergences of the two forecasts from the true distribution ?. Reference-based normalizations of the ignorance score are also possible, however, for the specific purpose of directly comparing M_1 and M_2 , we considered the additive difference more informative.

32. **Referee comment.** 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.

Response. We have replaced the unclear sentence by: “For the PSS, Eqs. C2 and C3 in Appendix C express the scores of M_1 and M_2 as expectations over the distribution of p_t , which can be evaluated numerically.”

33. **Referee comment.** Line 284: “In the regime $N \rightarrow \infty$, we obtain the the following asymptotic behavior for $p \rightarrow 0$.” Does this not also depend on the limiting behaviour of pN ?

Response. We agree with the reviewer. As explained in our response to Comment 30, the revised manuscript now explicitly states that the joint asymptotic regime requires $N \rightarrow \infty$, $p \rightarrow 0$, and $Np \rightarrow \infty$ for M_1 . The original sentence has been replaced accordingly.

34. **Referee comment.** 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.

Response. In order to clarify this point, in the revised Eqs. 28 and 29, the constants previously denoted by C_1 and C_2 have been renamed $D_{1,\eta}$ and $D_{2,\eta}$ (to avoid confusion with C_p). Both are positive and independent of p for fixed η but depend on the noise-to-signal ratio and contain the model-dependent local variance factors V_1 and V_2 , respectively. We have made these dependencies explicit in the revised manuscript. We also define $K_\eta > 0$ as

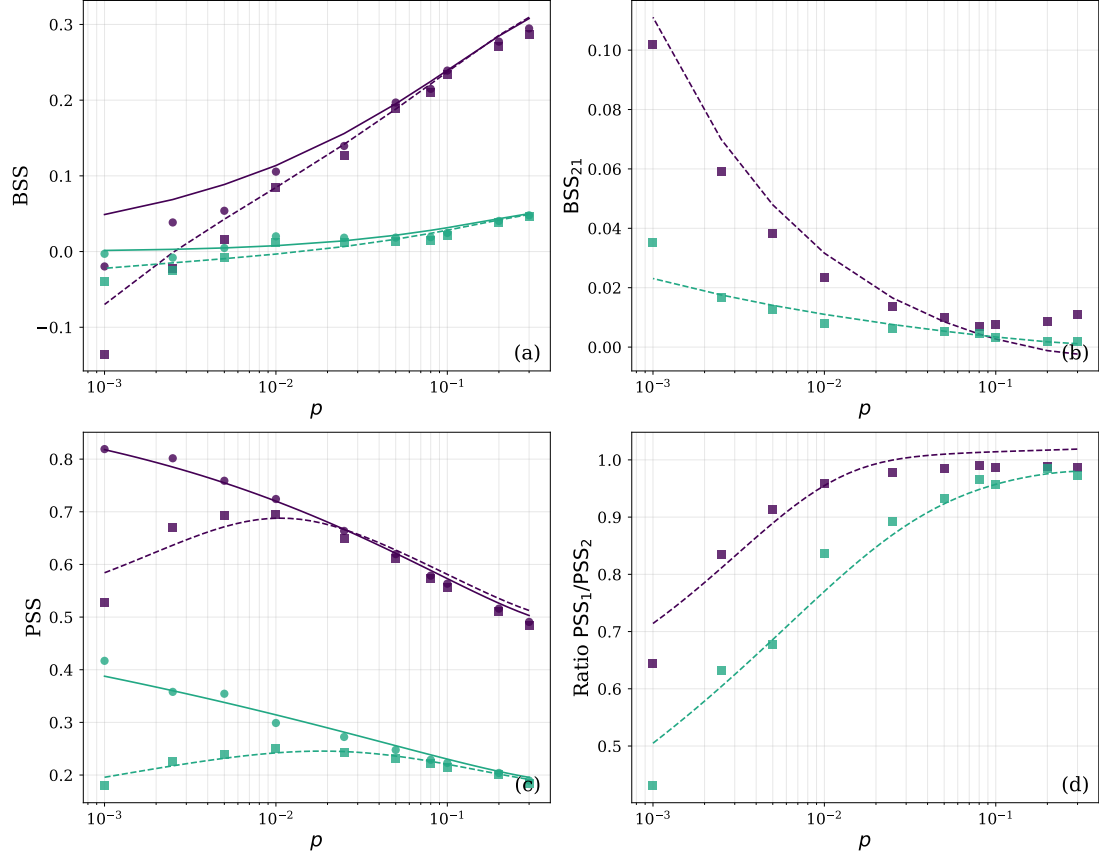


Figure 1: Comparison of Brier Skill Score (BSS) and Peirce Skill Score (PSS) for models $M_1 \in \mathcal{M}_1$ (symbols (■) and dashed lines) and $M_2 \in \mathcal{M}_2$ (symbols (●) and solid lines). Dark violet represent data for $\eta^2 = 1$ and while green represent data for $\eta^2 = 10$. Panel (a) shows empirically estimated BSS a function of exceedance probability p and panel (c) represents for Brier Skill Score of method 2 as compared to Method 1. Panel (c) presents analogous PSS results and panel (d) illustrates the PSS performance ratios PSS_1/PSS_2 against p . Dashed and solid lines in all panels show theoretical predictions.

the coefficient of the leading rare-event correction to the PSS based on the true conditional probability. It depends on η but not on p .

35. **Referee comment.** 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.

Response. We agree that the previous notation did not clearly distinguish the different quantities. We have revised it throughout the manuscript. The quantities $K_1(\eta)$ and $K_2(\eta)$ now denote exclusively the prefactors in the asymptotic prediction errors of M_1 and M_2 , respectively. The coefficient K_η denotes the leading rare-event correction to the PSS based on the true conditional probability while $D_{1,\eta}$ and $D_{2,\eta}$ denote the estimation-error corrections for the two methods. Finally, the coefficient in the PSS ratio is now denoted by $A_{p,\eta}$ to make its dependence on both p and η explicit. The former notation K_p has been removed and these quantities are now defined when first introduced.

36. **Referee comment.** Line 312: What is meant exactly by the term “logits loss”?

Response. Corrected, the term was actually referring to the sigmoid activation function and the term “loss” was not appropriate.

37. **Referee comment.** Line 316: Why is $d = 12$ chosen? Does this choice have any impact on the results?

Response. The value $d = 12$ was chosen as a compromise between the accuracy of the Gaussian approximation for the latent signal, which is a sum of d independent contributions, and the number of model parameters and the associated computational cost. To assess the sensitivity of the results to this choice, we repeated the experiments with $d = 10$ and $d = 18$. These values produced similar error and skill-score behaviors, including the relative performance of M_1 and M_2 . We have added this clarification to the revised manuscript.

38. **Referee comment.** 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 .

Response. Indeed, this is a good idea. We have followed reviewer’s suggestion and changed Figure 1 accordingly in the revised version.

39. **Referee comment.** 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?

Response. We agree that our formulation was insufficiently precise. For each test covariate, we estimated the bias as the difference between the mean prediction over repeated training samples and the true conditional probability p_t . The squared bias and the prediction variance were then averaged over the test covariates. For the configurations and exceedance probabilities shown in Fig. 1, the integrated squared-bias contribution was approximately 5 to 10 times smaller than the integrated variance. We have clarified this calculation in the revised manuscript and replaced the stronger term “validating” by “supporting” the variance-dominated approximation.

40. **Referee comment.** 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.

Response. We agree that the connection between the curves, the finite- p expressions in Appendix A, and Eqs. 24 and 25 was not sufficiently clear. The revised text now states that the curves in Fig. 1 show the finite- p approximations given in Eqs. 22 and 23 from which the asymptotic rare-event expressions in Eqs. 24 and 25 are derived. We also clarify that the model-dependent local variance factors V_1 and V_2 are fitted separately for each method and each value of η^2 . Consequently, the figure tests the predicted dependence on p rather than providing a parameter-free prediction of the absolute error amplitudes.

41. **Referee comment.** 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.

Response. We agree that the previous reference to a “quadratic correction” was unclear and could suggest that an empirical correction to the theoretical scaling was required. As

explained in our response to Major Comment 3, this term and the corresponding polynomial representation have been removed from Sect. 3 and from the revised figures. After retaining the corrected finite- p Fisher factor $C_p = p - \Phi_2(\Phi^{-1}(p); r)$ and repeating the simulations with twice the original sample size, the empirical dependence on p is satisfactorily reproduced using a single constant local variance factor V_k for each curve. More general covariate-dependent local variance functions, including polynomial forms, are now mentioned in Appendix A only as possible analytical extensions and are not used in the reported results.

42. **Referee comment.** 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.

Response. It is clear that the previous statement was somehow too strong and did not identify the assumptions being assessed. We have removed the claim of “excellent concordance” and “empirical validation.” The revised text now states explicitly that the analytical curves rely on large-sample asymptotic normality, negligible squared bias, the Fisher-information factorization for M_1 and approximately constant local variance factors. Since one model-dependent variance factor is fitted for each curve, the comparison primarily assesses the predicted dependence on p , rather than validating all assumptions or providing a parameter-free prediction of the absolute error amplitude. We now formulate more cautiously that the overall agreement supports the proposed theoretical expressions and their associated rare-event scalings over the probability range considered.

43. **Referee comment.** 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 M_1 , which apply a sigmoid function to some mean prediction). How are the results expected to change in this case?

Response. We thank the reviewer for this important point. We first clarify that M_1 does not apply a sigmoid to an estimate of the conditional mean of the continuous response. It has the general form $\hat{p}_t^{(1)} = \text{Sig}\{L(X_t; \theta)\}$, where L is an arbitrary learned logit. A neural classifier with a scalar output before the final sigmoid therefore has this functional form. We nevertheless agree that the regular asymptotic, large-sample analysis should be interpreted as a theoretical reference framework rather than as an exact finite-sample theory for over-parametrized neural networks used in the applications. In practice, model complexity, regularization, early stopping and optimization variability may alter the error amplitudes, convergence rates and probability level at which data scarcity becomes dominant. These effects are expected to be particularly important for M_1 , because information about positive outcomes is concentrated in approximately Np training examples. Model M_2 is also affected by finite-sample and complexity effects, but its distributional parameters are learned from all N continuous responses. As mentioned before, we now present the theoretical rates as a regular benchmark that isolates this information-loss mechanism, rather than as universal finite-sample rates for neural-network estimators.

44. **Referee comment.** 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?

Response. We agree that the previous sentence was ambiguous. What we intended to state is that the first processing stage only reorganizes the raw MeteoNet database into a

station-centered format, without applying filtering, imputation, or quality-control criteria. Checks for missing values, temporal completeness, and data consistency are instead performed in the subsequent processing stage. We have rewritten the corresponding paragraph to make this distinction explicit.

45. **Referee comment.** 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?

Response. We agree that the expression “to ensure statistical significance” was inappropriate and that the meaning of the thresholds was not clear. The values 2000 and 500 are not significance thresholds; they are minimum data-availability criteria applied to the number of valid observations retained after quality control for wind speed and accumulated rainfall, respectively. Their purpose is to exclude stations characterized by excessively short or incomplete time series while retaining a sufficiently large spatial sample of stations. We have rewritten the paragraph accordingly and removed the previously undefined term “keys”.

46. **Referee comment.** 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 M_1 and M_2 .

Response. Improved. The role of the different data sources and their relation to M_1 and M_2 was not made sufficiently explicit at the beginning of Section 4. We have therefore added a short overview clarifying that recent station observations and AROME/ARPEGE NWP forecasts are used as predictors by the same hybrid neural-network architecture, while future station observations define the prediction targets. The two strategies differ in their output and training objective: M_1 directly predicts threshold-exceedance probabilities and is trained separately for each threshold, whereas M_2 predicts the parameters of a conditional distribution from which exceedance probabilities are subsequently derived. Detailed descriptions of the inputs and architecture are retained in the following subsections.

47. **Referee comment.** 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?

Response. We have clarified both points in the revised manuscript. The spatial grid now refers explicitly to a local NWP subdomain centered on the ground station of interest. The meaning of the NWP data dimensions and the procedure used to extract these station-centered subgrids are described more explicitly in the text and in the table caption. For ARPEGE, the additional dimension corresponds to the vertical structure of the forecast fields, which are available at multiple atmospheric levels; these levels are retained as an additional input dimension in the model.

48. **Referee comment.** 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 $F_{C,+}^S(Q_p) = p_{+,R}$ should be $F_{C,+}^S(Q_p) = 1 - p_{+,R}$? 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?

Response. We thank the reviewer for identifying these notation inconsistencies. We now use $\Pr$ for probability operators and reserve $P(x)$ for the conditional-probability function defined explicitly in Eq. (17) and Appendix C. The quantile relation has also been corrected to $F_{C,+}^S(Q_{p,S}) = 1 - p_+$ where the superscript S identifies the station and the symbol $+$ indicates conditioning on positive rainfall. The former subscripts R and W , which distinguished rainfall and wind speed, respectively, have been removed to simplify the notation. These conventions are now used consistently throughout the revised manuscript.

49. **Referee comment.** Line 409: The probability $\bar{p}_{wet}$ does not have a bar in Appendix E (e.g. equation E2).

Response. A misprint has been corrected and it has been clarified that such notation is intentional. In Appendix E, p_{wet} refers to the wet-event probability for a single site. In Section 4, however, rainfall metrics are aggregated over many stations, each characterized by a different wet-event probability p_{wet} . We therefore use $\bar{p}_{wet}$ to denote the average rain probability across the considered stations when mapping the conditional rainfall probability to an unconditional probability axis p common to all stations. This averaging introduces a small approximation in the graphical representation, since the actual wet-event probability varies from station to station. We have clarified this distinction in the text.

50. **Referee comment.** Line 426: “The parameter ν controls the mean level.” This parameter is itself determined by two other parameters. Why are these not mentioned here

Response. The reviewer is right and we have clarified that ν is the non-centrality parameter defined by

$$\nu = \sqrt{\mu_1^2 + \mu_2^2},$$

where μ_1 and μ_2 are the means of the two components of the underlying bivariate conditional Gaussian distribution. Thus, the M-Rice distribution depends on the mean vector only through its Euclidean norm ν .

51. **Referee comment.** 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?

Response. Yes, and we rephrased the sentence so that now there is no more ambiguity on that matter.

52. **Referee comment.** Line 451: I assume D_S should be D_k ?

Response. Yes, corrected

53. **Referee comment.** 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?

Response. The number of input variable is the same in every site, but the number of observation and NWP variables chosen to predict wind and accumulated rainfall is not the same, so that n_S is notation for "number of ground observation predictors", and similarly n_{AR} indicates the number of used AROME NWP predictors. We slightly rephrased the text to introduce such notation more clearly.

54. **Referee comment.** Line 466: The notation $AR_{d,t}^S$ and $AP_{d,t}^S$ differs from that employed previously. I assume these should be AR_k and AP_k ?

Response. Yes, this has been corrected.

55. **Referee comment.** 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.

Response. Yes we meant the same specific date and not the same calendar day, this error in the text has been corrected.

56. **Referee comment.** 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?

Response. We agree that the percentage of dates assigned to the test subset may appear small. However, each selected date contains hourly samples of the order of several hundred stations, so the resulting evaluation sample is substantially larger than the percentage of dates alone suggests. The test subsets contain approximately 2×10^5 valid samples for each predictand, while the combined validation and test subsets contain approximately 6×10^5 . Even at the rarest probability level considered, the test subset therefore contains several hundred observed events. The scores were also computed separately on the validation and test subsets, and M_2 outperformed M_1 in both cases. We have not performed a formal uncertainty analysis of the score differences, so we do not attach confidence intervals or significance levels to individual score differences. Our conclusions are instead based on the consistency of the observed pattern across thresholds, forecast horizons, metrics, and the two evaluation subsets. We have added the effective sample sizes to the revised manuscript.

57. **Referee comment.** Line 486: “As reported in Table 1, ...” Information about the total number of keys is not reported in Table 1.

Response. Yes this text has been changed. We thank the reviewer for having pointed it out

58. **Referee comment.** Line 525: “the term Rel_U , even if increasing rapidly with $p \rightarrow 0$, remains relatively well controlled.” What is meant by the term “well controlled”?

Response. We have checked the evolution of the components $\frac{\text{Rel}}{U}$ and $\frac{\text{Res}}{U}$ with $p \rightarrow 0$, for the two strategies considered and for both wind and rainfall, $\frac{\text{Rel}}{U}$ remains smaller than the corresponding resolution contribution over the considered range even if we observe a worsening with decreasing p . This implies that BSS is always positive. We are now more precise in the text.

59. **Referee comment.** 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.

Response. We agree that a climatology-referenced skill measure can also be constructed from the logarithmic score. In the revised manuscript, however, we use the BSS as the primary climatology-referenced measure of probabilistic skill, since this formulation is well established in meteorological forecast verification. The LS is retained only as a complementary proper scoring rule and is therefore reported for a limited set of representative cases rather than displayed as a function of the event base rate. Similarly, the AUC is included as a complementary threshold-independent measure of discrimination. The Table format also

allows to present results in a more compact format than figures, so additional skill measures can be presented without exceeding maximum text length.

60. **Referee comment.** 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.

Response. We agree, and these two complementary skill measures have been moved in a dedicated section in the Appendix.

61. **Referee comment.** 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.

Response. We agree that this notation was not useful and confusing. It has been removed in the revised version of the manuscript.

62. **Referee comment.** 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.

Response. We agree that the previous discussion incorrectly described aggregated proper scores as an “operational proxy for reality” and could suggest a direct correspondence between the selected model family and the true conditional distribution. We have substantially rewritten this subsection. The revised text now presents strictly proper scoring rules as principled measures of out-of-sample predictive performance, rather than as proxies for the true conditional law. It also clarifies that, in finite samples, several candidate families may yield similar average scores and that, under model misspecification, their ranking may depend on the selected scoring rule and prediction task. Similar scores therefore indicate limited empirical distinguishability, rather than structural equivalence between the candidate families. Finally, we have moderated the physical interpretation. The revised text states that the limited sensitivity observed for the reasonably suitable families examined here is consistent with a substantial part of the predictive information being associated with changes in bulk distributional properties. This interpretation is restricted to the present datasets and candidate families and is not presented as a universal property of distributional forecasting.

63. **Referee comment.** Line 599: These results hold for all metrics considered, not just when predictive skill is “measured by the PSS”.

Response. We agree that M_2 outperforms M_1 for all the metrics considered. However, the scores do not have the same dependence on the event rarity: the positively oriented PSS increases toward rarer events, whereas the BSS generally decreases, as shown in Fig. 2. We have revised the text to distinguish the relative ranking of the two methods from the evolution of each score with p .

64. **Referee comment.** 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 M_2 would generally perform worse than M_1 . I do not agree that this is improvement is “intrinsic” in all distributional models.

Response. We agree that the previous statement was too general and too strong. We have revised the conclusion to clarify that the limited sensitivity observed here applies only to the reasonably suitable parametric families. It should not be interpreted as an intrinsic robustness of distributional models to arbitrary misspecification. We now explicitly acknowledge, as suggested by reviewer, that an inappropriate family may substantially degrade the performance of M_2 and remove its advantage over direct classification.

65. **Referee comment.** Line 620: What is meant by a “continuous spatial domain”? All spatial domains considered in practice are in some sense discretized.

Response. We agree that the expression “continuous spatial domain” was imprecise. Since the Conclusions have been rewritten to focus on directions more directly related to the scope of the present study, the sentence has been removed and spatial extensions are no longer discussed.

Appendix corrections

Referee comment. Appendices: The notation used throughout the appendices (particularly Appendix A) is quite inconsistent, making the theoretical results difficult to follow. For example:

- **Referee comment.** 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?

Response. In Sec. 2, we have reorganized the definition of class $\mathcal{M}_2$ to introduce first the chosen parametric density $f(y; \Pi)$ and the estimated parameter vector $\hat{\Pi}_t = M_2(X_t; \hat{\theta})$. We then state explicitly that, in the Gaussian location model of Sect. 3, $\Pi_t = \mu_t$ and $M_2(X_t; \hat{\theta}) = \hat{\mu}_t$, where $\hat{\mu}_t$ is the estimated conditional mean. In Sec. 3, we now have the sentence: “We also consider a model $M_2(\cdot, \theta)$ in the class $\mathcal{M}_2$ which provides an estimation of μ_t , denoted as $\hat{\mu}_t$, allowing...”. The corrected the misprint, it was X instead of Z .

- **Referee comment.** 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.

Response. Reviewer is right, this is now stated explicitly in Appendix A.

- **Referee comment.** Line 648: $\nabla_{\theta}^2 \ell(Y, X; \theta_0)$ should be $\nabla_{\theta}^2 \ell(Y, X; \theta)|_{\theta=\theta_0}$. Similarly for the definition of $G(X)$ on line 651.

Response. We agree and have replaced both expressions by explicit evaluations at $\theta = \theta_0$.

- **Referee comment.** $\hat{M}_1(X_t; \theta)$, $\hat{M}_2(X_t; \theta)$, $\hat{z}(X_t; \theta)$ are sometimes written with a semi-colon and sometimes with a comma. The dependence on θ is also sometimes suppressed, but not always.

Response. We have harmonized the notation throughout the manuscript. A semicolon is now used systematically to separate covariates from model parameters and the parameter dependence is omitted only after evaluation at the estimated or population parameter when no ambiguity can arise.

- **Referee comment.** Equation A6: What exactly does $\approx$ mean here? The symbol $\simeq$ 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.

Response. We agree that the notation was insufficiently explicit. In the revised Appendix A, $\approx$ denotes a first-order delta-method approximation and $\simeq$ denotes an additional finite- p modelling approximation/assumption such as the Fisher-information factorization or the constant-local-variance assumption. The symbol $\sim$ denotes asymptotic equivalence, meaning that the ratio of the two quantities tends to one in the specified limit. These conventions are now stated at the beginning of Appendix A. We have also rewritten Eq. A6 with an explicit $o(N^{-1})$ remainder and clarified that the variance is taken over repeated training samples, conditionally on X_t .

- **Referee comment.** 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 ?

Response. Indeed, our previous notation was ambiguous. In the revised manuscript, we introduce a deterministic argument u and define

$$\bar{V}_2(u) = \mathbb{E}[V_2(X_t) \mid \mu(X_t) = u].$$

This is a version of the conditional expectation as a function of the value taken by $\mu(X_t)$, and the notation has been revised accordingly. To simplify the notation, all quantities previously denoted by V'_k are now denoted by V_k , since no corresponding unprimed quantities were used elsewhere.

- **Referee comment.** 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.

Response. We agree. The revised notation now distinguishes the original local variance $V_2(X_t)$ from its conditional average

$$\bar{V}_2(u) = \mathbb{E}[V_2(X_t) \mid \mu(X_t) = u].$$

The integral is now written in terms of $\bar{V}_2(u)$.

- **Referee comment.** 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 $Q_p = -(\sigma/\rho)\sqrt{1 + \rho^2} \Phi^{-1}(p)$. 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?

Response. This was indeed an ambiguity in first manuscript version. In the previous expression, the integration variable was implicitly standardized as $u = \mu/s$, but this change

of variable was not stated. We have rewritten the integral using the original unstandardized variable, so that all scale factors are now explicit and no factor is implicitly extracted from $(Q_p - u)^2$.

- **Referee comment.** 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.

Response. This expression has been removed in the revised version and the approximation of $\bar{V}_2(u)$ is now well defined in Eq. (A7).

- **Referee comment.** Equation A9: V_2 is defined as a function but is here treated as a constant?

Response. We agree that, here again, the previous notation was ambiguous. As previously explained, in the revised manuscript, $V_2(X_t)$ denotes the original local variance and $\bar{V}_2(u) = \mathbb{E}[V_2(X_t) \mid \mu(X_t) = u]$ its conditional average. Equation A7 then explicitly introduces the approximation $\bar{V}_2(u) \simeq V_2$, where the same symbol without an argument denotes the positive constant approximation. This convention was chosen to avoid introducing an additional notation and has now been stated explicitly.

- **Referee comment.** Equation A16: What is $\hat{p}^{(3)}$?

Response. This was a typographical error and has been corrected to $\hat{p}^{(1)}$.

- **Referee comment.** Appendix C: Many of the integrals in this section are missing dx .

Response. This has been corrected.

- **Referee comment.** Line 730: The notation $\mathbb{E}[I_{t+h} \mid p_t]$ has not been defined, and is different from the notation for conditional expectations and probabilities elsewhere in the paper (e.g. line 665).

Response. We agree that the notation was not introduced explicitly. We now state that, since $p_t = \Pr(I_{t+h} = 1 \mid X_t)$, the tower property gives $\mathbb{E}[I_{t+h} \mid p_t] = p_t$. Moreover, we have also harmonized the notation for conditional expectations throughout the revised manuscript.

- **Referee comment.** 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.

Response. We agree with the reviewer. The previous statement was a consequence of consistency rather than its standard definition, and the notation δ_t also conflicted with that used for the Dirac distribution. We have replaced it with the standard consistency condition $\hat{p}_t^{(k)} \xrightarrow{p} p_t$ as $N \rightarrow \infty$. We also state explicitly that the variance-only expression obtained from the second-order Taylor expansion relies on a negligible-bias approximation.

- **Referee comment.** 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.

Response. Indeed, our notation was unclear. In the revised manuscript, we now define $P(x)$ unambiguously as $P(x) = \Pr(I_{t+h} = 1 \mid X_t = x)$ so that $P(X_t) = p_t$.

- **Referee comment.** 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.

Response. Yes, the symbol Z was inconsistent with the notation used elsewhere and has been replaced by X . We have also corrected the dimension: each covariate vector satisfies $X_t \in \mathbb{R}^d$ while the complete sample $(X_1, \dots, X_N)^\top$ can be represented as an $N \times d$ matrix.

- **Referee comment.** 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.

Response. We agree that the notation was insufficiently explained. Here n denotes the order of the Gauss-Hermite quadrature and is unrelated to the sample size N used elsewhere in the manuscript; H_{n-1} is the Hermite polynomial of order $n - 1$, and y_i are the n roots of H_n , i.e. the Gauss-Hermite quadrature nodes. We have clarified these definitions in the revised text. We also clarify that Eq. (E1) is not a sample estimate of the mean, but a deterministic numerical approximation of its integral representation obtained using Gauss-Hermite quadrature.

- **Referee comment.** Appendix E2: The parameter vector θ is not bold in this section, in contrast to in all other sections.

Response. This was another inconsistency in the notation, as θ was used to denote the learnable parameters of the NN throughout the text, while in this appendix θ stands for the parameters of the chosen pdf form. We now denote the latter with Π . We thank the reviewer for pointing this out.

References

- Bentzien, S. and Friederichs, P.: Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE, *Weather and Forecasting*, 27, 988–1002, 2012.
- Primo, C., Schulz, B., Lerch, S., and Hess, R.: Comparison of model output statistics and neural networks to postprocess wind gusts, in: *Symposium on Applications of Mathematical Sciences*, pp. 153–180, Springer, 2023.
- Weijs, S. V., Van Nooijen, R., and Van De Giesen, N.: Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition, *Monthly Weather Review*, 138, 3387–3399, 2010.
- Williams, R., Ferro, C., and Kwasniok, F.: A comparison of ensemble post-processing methods for extreme events, *Quarterly Journal of the Royal Meteorological Society*, 140, 1112–1120, 2014.