

Complete Response to the Referee 3

Manuscript egusphere-2026-3111

Roberta Baggio and Jean-François Muzy

1 Reviewer comments

1.1 Major comments

Referee comment. The manuscript by Baggio and Muzy contributes to the important field of building tools for the forecast of severe weather, which is relevant for disasters anticipation and mitigation. They evaluated direct binary classification and full-distribution parametric modeling, and validate their methods in a Mediterranean region of southeastern France for strong near-surface winds and heavy rainfall using ground-station data from Météo-France.

In my view, the manuscript is well organized, uses a correct technical terminology, and presents a research topic that is relevant. However, I found the text difficult to read and understand by a wide and general audience. Although correct, dense mathematical notation and frequent specific terminology makes it complex and difficult to follow. For example, in the results section, despite of the use of ground-based variables that are commonly used for describing meteorological conditions (hourly wind speed and accumulated rainfall), it is somewhat confusing to interpret and connect the Brier and Peirce skill scores to natural weather likelihood.

The manuscript is valuable, but in the current format it is dense and hard to follow, even for researchers in Meteorology and Climate science with good mathematical and statistical training. Without a review in its overall style, I believe the manuscript is more suitable for academic journals in Applied Mathematics or Machine Learning.

Response. We thank the reviewer for this valuable feedback. Following the reviewer's suggestions, we have revised the manuscript in several places. We believe that the revised version is now clearer and more readable, that the novel aspects of the study are stated more explicitly, and that the main results are presented and interpreted more clearly. Among the changes introduced, we would particularly like to highlight the following:

1. **The role of the theoretical analysis in Section 3 has been clarified and made more explicit:**

In both the Introduction and the opening of Section 3, we now state more clearly the scope and purpose of the simplified analytical framework. The theoretical model is not intended to provide a direct representation of the behavior of real meteorological variables. Rather, it provides a controlled setting in which the statistical consequences of replacing a continuous predictand with a thresholded binary outcome can be characterized explicitly.

In particular, thresholding the predictand entails a loss of information for the binary-classification approach, which leads to a different dependence of the probability-estimation error on the event probability p compared with the full-distribution approach. This information loss also propagates to the verification metrics considered in the study, which were selected to provide complementary assessments of discrimination and probabilistic forecast quality in the rare-event regime.

We now also make clearer in the Introduction that comparisons between direct binary classification and full-distribution modeling have already been considered in several previous

studies. The novelty of our contribution is therefore not the comparison itself, but rather the fact that we carry it out within a deliberately simplified and controlled framework, using the same modeling strategy for both approaches and a tractable underlying statistical model. This setting allows us to derive explicit asymptotic scalings for the relative estimation errors and for the associated verification scores as $p \rightarrow 0$.

These analytical results then provide an interpretive framework for the applications to real meteorological variables: rather than expecting the exact theoretical scalings to hold quantitatively, we use the theoretical behavior to identify qualitative signatures that can be examined in the wind-speed and rainfall applications.

2. The meaning of the selected verification metrics, as well as the rationale for their choice, has been clarified:

We have substantially expanded Subsection 2.3 to better explain the meaning of the selected verification metrics and their relevance to the forecasting of increasingly rare events. Concerning deterministic metrics, we selected the Peirce Skill Score (PSS) as our primary measure of discrimination among all the possible categorical skill scores. The PSS is equitable and, because the hit and false-alarm rates are normalized separately over observed events and non-events, it has no explicit dependence on the event base rate. This makes it particularly suitable for comparing discrimination across events of strongly differing rarity. For probabilistic metrics, we consider the Brier Score (BS) and the Logarithmic Score (LS), both of which are strictly proper scoring rules and therefore encourage honest probability forecasts. We additionally use the Brier Skill Score (BSS), which expresses Brier-score performance relative to the climatological forecast. This normalization gives the score a direct interpretation in terms of improvement over climatology and is particularly useful when comparing events with very different base rates. Finally, we report the Area Under the ROC Curve (AUC) as a complementary threshold-independent measure of discrimination, summarizing the ability of the probabilistic forecasts to rank events above non-events over all possible decision thresholds. We hope that this more comprehensive motivation for the selected verification metrics, together with a more careful presentation and discussion of the results, facilitates a clearer interpretation of the study and of its main findings.

3. A brief summary of the analytical results and their implication has been added:

In line with the reviewer’s comment regarding the readability of the manuscript and the heaviness of the mathematical notation, we have streamlined the analytical development in Section 3 and moved most of the technical derivations and intermediate expressions to the Appendices. At the same time, a certain amount of analytical detail necessarily remains in the main text, as the derivation and interpretation of the rare-event asymptotic behavior constitute a central component of the proposed study.

To further improve readability, we have added a new subsection at the end of Section 3 summarizing and interpreting the main theoretical results. In particular, we emphasize that the simplified framework makes it possible to derive explicitly the relative advantage of the full-distribution approach over direct binary classification and to characterize how this advantage evolves as the event probability p decreases. We also clarify which aspects of the resulting behavior are specific to the adopted Gaussian location model. In particular, the increase in discrimination, as measured by the PSS, toward more extreme events arises from the specific mean-shift mechanism of this model: increasingly rare exceedances become progressively associated with large excursions of the predictable location component.

This distinction also provides a clearer motivation for the meteorological applications presented in the following section. Rather than expecting the exact theoretical behavior to hold for real atmospheric variables, we examine which qualitative features of the simplified framework persist in the wind-speed and rainfall applications. Differences between the theoretical and

empirical behavior can therefore also provide useful insight into the statistical mechanisms underlying the predictability of extreme events.

4. **The Conclusions have been revised to highlight the main findings more clearly.**

We have revised the Conclusions to provide a clearer and more explicit summary of the main contributions of the study. In particular, we now emphasize the twofold purpose of our analysis.

First, the study provides a direct comparison of two paradigms for threshold-exceedance forecasting within a controlled and analytically tractable setting. This framework makes it possible to characterize explicitly the information loss induced by thresholding the continuous predictand and to quantify how its consequences become increasingly important as the event probability p decreases.

Second, we emphasize that the similarities and differences between the analytical results and the real meteorological applications are themselves informative. Both meteorological variables considered in the study exhibit the increasing relative disadvantage of direct binary classification toward rarer events. Wind speed shows the closest qualitative correspondence with the simplified theoretical framework. This is not because wind speed follows the Gaussian assumptions of the toy model, which it clearly does not, but because its discrimination increases toward rarer thresholds, consistently with a mechanism in which increasingly extreme events become progressively associated with stronger predictable changes in the bulk of the conditional distribution.

Rainfall exhibits a somewhat different behavior. Although several qualitative trends predicted by the simplified framework remain visible, the discrimination of the full-distribution approach does not increase appreciably toward the rarest thresholds considered. Nevertheless, the full-distribution approach consistently outperforms direct binary classification, and its relative advantage increases toward rarer events. This illustrates how the information-loss mechanism identified analytically may remain relevant even when the specific rare-event behavior of the Gaussian location model does not carry over directly to the application.

Finally, we have clarified the main directions for future work, including a more systematic investigation of the consequences of distributional misspecification and extensions of the neural-network architecture used in the present study.

1.2 Minor comment

Referee comment. Line 505: Please correct the word “Applicaton” in “4.4 Applicaton results”, which I believe is “Application”.

Response. We have corrected this error as many other typos we found.