arXiv:1804.09233v2 [stat.AP] 5 Mar 2019

Post-processing multi-ensemble temperature and precipitation forecasts through an Exchangeable Gamma Normal model and its Tobit extension

Marie Courbariaux[1] Pierre Barbillon[1*] Luc Perreault[2]

Éric Parent[1]

[1]UMR MIA-Paris, AgroParisTech, INRA,
Université Paris-Saclay, 75005, Paris, France
[2]Hydro-Québec Research Institute, Varennes, Canada

March 7, 2019

Abstract

Meteorological ensemble members are a collection of scenarios for future weather issued by a meteorological center. Such ensembles nowadays form the main source of valuable information for probabilistic forecasting which aims at producing a predictive probability distribution of the quantity of interest instead of a single best guess point-wise estimate. Unfortunately, ensemble members cannot generally be considered as a sample from such a predictive probability distribution without a preliminary post-processing treatment to re-calibrate the ensemble. Two main families of post-processing methods, either competing such as the BMA or collaborative such as the EMOS, can be found in the literature. This paper proposes a mixed effect model belonging to the collaborative family. The structure of the model is formally justified by Bruno de Finetti's representation theorem which shows how to construct operational statistical models of ensemble based on judgments of invariance under the relabeling of the members. Its interesting specificities are as follows: 1) exchangeability contributes to parsimony, with an interpretation of the latent pivot of the ensemble in terms of a statistical synthesis of the essential meteorological features of the ensemble members, 2) a multi-ensemble implementation is straightforward, allowing to take advantage of various information so as to increase the sharpness of the forecasting procedure. Focus is cast onto Normal statistical structures, first with a direct application for temperatures, then with its very convenient Tobit extension for precipitation. Inference is performed by Expectation Maximization (EM) algorithms with both steps leading to explicit analytic expressions in the Gaussian temperature case and recourse is made to stochastic conditional simulations in the zero-inflated precipitation case. After checking its good behavior on artificial data, the proposed post-processing technique is applied to temperature and precipitation ensemble forecasts produced for lead times from 1 to 9 days over five river basins managed by Hydro-Québec, which ranks among the world's largest electric companies. These ensemble forecasts, provided by three meteorological global forecast centres (Canadian, US and European), were extracted from the THORPEX Interactive Grand Global Ensemble (TIGGE) database. The results indicate that post-processed ensembles are calibrated and generally sharper than the raw ensembles for the five watersheds under study.

KEYWORDS: Hierarchical latent variable models, EM algorithms, Ensemble numerical weather prediction, Statistical post-processing, Temperature, Precipitation

1 Introduction

Rather than a single scenario (*i.e. a deterministic prediction*), meteorological services are now delivering a full collection of scenarios (that are called *the members of the ensemble*) as an attempt to depict their knowledge as well as their uncertainty about the future state of the atmosphere. The members of the ensemble are obtained by introducing perturbations into the initial conditions or the parametrization of a numerical weather prediction model (NWP), or by using several NWP models. As an example,

*corresponding author: pierre.barbillon@agroparistech.fr

the European Center for Medium-range Weather Forecasts (ECMWF) generates the 50 members of the ECMWF-EPS (Ensemble Prediction System) by launching every day 50 runs of their NWP model with different initial conditions as described in [Buizza et al., 2008].

One may expect that the ensemble members $X_{k,t,h}, k = 1 \dots K$, issued at time $t - h$, will behave as a K-sample from the probabilistic forecast of the variable Y_t to be predicted h days ahead (e.g. temperature, wind, precipitation and so on). This hypothesis of *calibration* or *reliability* for meteorological ensembles is sometimes assumed by weather forecasts users such as hydropower companies. However this is generally not tenable for most ensemble prediction systems. In fact, as pointed out by, among others, Hamill and Colucci [1997], Bougeault et al. [2010], Taillardat et al. [2016] and Perreault [2017], ensemble weather forecasts are subject to bias and tend to be underdispersive. In the following, the index h will be omitted and we will write $X_{k,t}$ instead of $X_{k,t,h}$ since the lead time h is considered fixed.

As an illustration, the three top panels of Figure 1 show the rank histograms for 3-days ahead precipitation ensemble forecasts produced in 2014 for Manic 2 watershed by the European (ECMWF-EPS), Canadian (CMC[1]-EPS) and US (NCEP-GEFS[2]) meteorological services. The marked deviations from the uniform distribution of these histograms show that these ensembles are all far from being calibrated: the raw ensembles are biased and under-dispersed [Hamill, 2001]. It is therefore necessary to carry out some form of statistical post-processing on the ensemble members in order to simulate a new K-sample, which will then be well calibrated. Several methods of statistical post-processing have been proposed to improve ensemble weather forecasts. Two main strategies can be found in the literature: the members will either compete or collaborate.

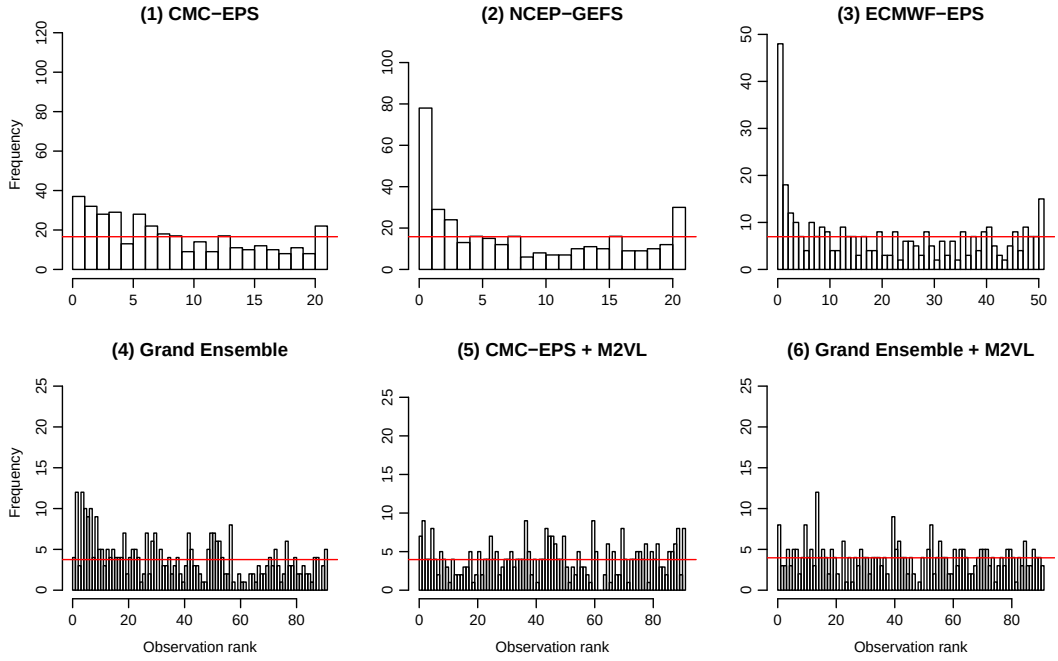


Figure 1: Rank histograms for 3-days lead time ensemble precipitation forecasts produced daily in 2014 for Manic 2 watershed: (1) CMC-EPS raw ensemble (2) NCEP-GEFS raw ensemble (3) ECMWF-EPS raw ensemble (4) large raw ensemble gathering CMC-EPS, NCEP-GEFS and ECMWF-EPS members (5) post-processed CMC-EPS (6) post-processed large ensemble.

The first family considers that each member k in itself can be the potential '*truth*' $X_{k,t}$ for the quantity Y_t to forecast. As if the K members of the ensemble were competing, the idea is to identify statistically some *best* member k^* and try to generate ensemble forecasts in the vicinity of $X_{k^*,t}$. To provide an adjusted predictive distribution, one generally relies on weighted smoothing kernels over the ensemble members. The statistical structure underpinning most methods from this family is as follows:

[1]CMC: Canadian Meteorological Center.
[2]NCEP: National Center for Environmental Prediction; GEFS: Global Ensemble Forecast System.

for any t and k,

$$\begin{aligned}(Y_t|\mathbf{X}_t, k^\star) &\sim \mathcal{G}\left(X_{k^\star,t}, \boldsymbol{\theta}_{k^\star}\right) \\ Pr(k^\star = k) &= \pi_k\end{aligned} \tag{1}$$

where $\mathbf{X}_t = \{X_{1,t}, \ldots, X_{K,t}\}$, $\sum_{k=1}^{K} \pi_k = 1$, $\mathcal{G}$ is a probability distribution function (pdf) to be chosen, $\boldsymbol{\theta}_k$ and π_k, for $k \in \{1, \ldots, K\}$, are parameters to be estimated and $k^\star$ is the index of the (unknown) best member. As it is usually done, post-processing is applied independently for each lead time, the temporal dependency being reconstructed using empirical copulas. The most famous method in the competing family is the Bayesian Model Averaging (BMA) proposed by Raftery et al. [2005]. They considered the Gaussian case, taking $\mathcal{G}$ under the following form: for any t and k,

$$\mathcal{G}\left(X_{k,t}, \boldsymbol{\theta}_k\right) = \mathcal{N}\left(a_k X_{k,t} + b_k, \sigma^2\right),$$

with $\boldsymbol{\theta}_k = (a_k, b_k, \sigma)$ for $k \in \{1, \ldots, K\}$. The EM (Expectation-Maximization) algorithm is used for inference and the resulting predictive distribution in this case is a mixture of Gaussian distributions. Of course, this is a way of re-assembling the members to work together for prediction at the end, but this collaboration is performed with uneven weights, at least in principle. Consequently the BMA can also be understood as a post-processing *dressing* method with a Gaussian kernel. This method is not without drawback when the ensemble is over-dispersive as shown by a simulated data experiment conducted by Raftery et al. [2005]. However this case rarely happens in practice since ensemble weather prediction systems tend to produce overconfident forecasts: the resulting ensemble members are generally less dispersed than they should be. Many extensions of the BMA have been proposed, changing the method of inference or the distribution of the ensemble. As an example for precipitation Sloughter et al. [2007] proposed a power transformation (1/3) and the following Bernoulli-Gamma model for $\mathcal{G}$ so as to deal with the case of no rain: for any t,

$$\begin{cases} \text{logit}\left(Pr(Y_t^{\frac{1}{3}} = 0 | k^\star = k, X_{k,t})\right) & = d_0 + d_1 X_{k,t}^{\frac{1}{3}} + d_2 \mathbb{I}_{(X_{k,t}=0)} \\ \left(Y_t^{\frac{1}{3}} | k^\star = k, X_{k,t}, Y_t > 0\right) & \sim \Gamma\left(\alpha\left(X_{k,t}\right), \beta\left(X_{k,t}\right)\right) \\ \alpha\left(X_{k,t}\right) \text{ and } \beta\left(X_{k,t}\right) & \text{s. t. } \frac{\alpha(X_{k,t})}{\beta(X_{k,t})} = b_0 + b_1 X_{k,t}^{\frac{1}{3}} \\ & \text{and } \frac{\alpha(X_{k,t})}{\beta(X_{k,t})^2} = c_0 + c_1 X_{k,t} \\ Pr(k^\star = k) & = \frac{1}{K} \ \forall k \end{cases}$$

where $\mathbf{d} = (d_0, d_1, d_2)$, $\mathbf{b} = (b_0, b_1)$ and $\mathbf{c} = (c_0, c_1)$ are vectors for parameters to be estimated and $\Gamma(\alpha, \beta)$ stands for the gamma distribution with shape α and rate β.

The second family of post-processing techniques for meteorological ensembles considers that members are not alternative individualized proposals for the quantity to be predicted, but rather a collection of scenarios sharing common traits that are identified as summaries of the future state of the system to be predicted. The predictive distribution is to be based on these shared features. The Bayesian Processor of Output (BPO) [Krzysztofowicz and Maranzano, 2006] suggests to consider the joint distribution $(\mathbf{X}_t, Y_t)$ (after a normal quantile transform) in order to derive the conditional distribution of Y_t given $\mathbf{X}_t$. Without any additional hypothesis on the form of covariance between members, this model lacks parsimony. To our view, this is a major drawback when considering the limited sample size of historical data available to learn about the distribution of the ensemble $\mathbf{X}_t$. A first step towards parsimony would assume that the weighted mean of the ensemble is the sufficient statistic for some parametric modeling of the predictive distribution (generally, the ordinary mean is used). One can then imagine treating the ensemble mean as a deterministic forecast and consider a bivariate Gaussian model to link this statistic with the quantity to forecast Y_t. The EMOS (Ensemble Model Output Statistics) method proposed by Gneiting et al. [2005] additionally assumes that the ensemble dispersion also informs on the variability of the future meteorological state. The proposed predictive parametric model is as follows: for any t,

$$(Y_t|\mathbf{X}_t) = a + \mathbf{b}^T \mathbf{X}_t + \sqrt{c + d S_{\mathbf{X}_t}^2}\,\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, 1), \tag{2}$$

where a, $c > 0$, $d > 0$ and $\mathbf{b}$ are parameters to be estimated, and $S_{\mathbf{X}}$ denotes the standard deviation of the ensemble $\mathbf{X}$. Conversely to the BMA, EMOS cannot provide a multimodal predictive distribution. One can also encounter post-processing methods for ensemble precipitation forecasts in the collaborative family, such as the EMOS-like model proposed by Scheuerer [2014]: for any t and k,

$$\begin{cases} (Z_t|\mathbf{X}_t) & \sim GEV\left(\mu_{\mathbf{X}_t}, \gamma + \kappa MD_{\mathbf{X}_t}, \xi\right) \\ \mu_{\mathbf{X}_t} \text{ s.t. } \mathbb{E}\left(Z_t|\mathbf{X}_t\right) & = \alpha + \beta\bar{\mathbf{X}}_t + \frac{1}{K}\sum_{k=1}^{k=K} \mathbb{I}_{X_{k,t}=0} \\ MD_{\mathbf{X}_t} & = \frac{1}{K^2}\sum_{k,k'} |X_{k,t} - X_{k',t}| \\ (Y_t|Z_t) & = Z_t \mathbb{I}_{Z_t \geq 0} \end{cases}$$

where α, β, γ, κ et ξ are parameters to be estimated, $MD_{\mathbf{X}}$ is a measure of dispersion for the ensemble $\mathbf{X}$, Z_t is the latent non censored variable for the precipitation Y_t and $GEV(\mu, \sigma, \xi)$ is the generalized extreme value distribution with location parameter μ, dispersion parameter $\sigma > 0$ and shape parameter ξ.

Finally, although they do not provide an explicit predictive function but allow for estimation of the quantiles, methods closely related to quantile regression and nonparametric regression can also be considered to belong to the family of collaborative post-processing techniques. For instance, following the logistic regression works of Wilks [2009], Ben Bouallègue [2013] proposes to add an interaction term in this post-processing technique and Messner et al. [2014] developed a heteroscedastic version in order to use the information contained in the dispersion of the members. Taillardat et al. [2016] combined quantile regression with random forests (Meinshausen [2006]), for the post-processing of ensemble temperatures and wind speed predictions.

The purpose of this paper is to formalize and develop a new collaborative post-processing technique in the vein of EMOS but allowing to incorporate the information conveyed by multiple ensembles into the analysis. Such a collaborative post-processing approach avoids the main shortcomings of dressing methods (a poor adaptation to cases where the ensembles are over-dispersive) and of the non-parametric methods, which are not originally intended to issue complete predictive distributions.

Multi-model ensemble prediction is increasingly used since number of studies have demonstrated that these forecasts have higher prediction skill than that of an individual model (Tebaldi and Knutti [2007]). Such "grand ensemble" are usually considered for long term meteorological forecasting, namely for seasonal forecasts (Khajehei et al. [2018]), and for hydrological forecasting. Even though using raw multiple sources of meteorological forecasts may improve reliability, they still lack of calibration. Except for the BMA approach which belongs to the competing family (Fraley et al. [2010]), no parametric method formally addresses multi-ensemble forecasts post-processing. Especially, to the best of our knowledge, no collaborative statistical post-processing model using latent variables, such as the EMOS extension presented herein, have been proposed to explicitly calibrate multiple ensembles forecasts. This is confirmed by Li et al. [2017] in their recent review on statistical post-processing methods for hydrometeorological ensemble forecasting.

In this paper, we focus on medium range forecasting of daily temperature and precipitation that are of utmost interest for hydropower companies, since these two quantities are the main inputs of the rainfall-runoff model used to produce streamflow predictions (Guay et al. [2018], Courbariaux et al. [2017], Garçon [1996]). Strategic decisions are taken daily on the basis of these forecasts, namely to prevent flooding damages and to avoid operating losses. The availability of reliable temperature and precipitation probabilistic forecasts in this context is therefore a crucial issue.

Being explicitly underpinned by the theory of exchangeability, the proposed technique will benefit from a statistical property of symmetry (invariance by relabeling) that can be naturally expected from efficient ensemble simulators. The statistical framework of exchangeability is recalled in Section 2. Its main interest lies in providing a theoretical justification for the representation of the ensemble as a sample from a mixed effect model. Figuring out the conditioning latent process as some salient configuration of the ensemble simulator for the day to predict can help to understand the mixed model in meteorological terms. Section 3 develops exchangeable applications for the Gaussian case with a noticeable multidimensional mixed model that might improve forecasting sharpness by taking into account the multiple sources of information conveyed by different ensemble simulators or by forecasts issued from various meteorological experts. This first model is suitable for temperature ensemble forecasts. Section 4 is devoted to the Tobit extension of the exchangeable Gaussian model to deal with the zero inflation of the precipitation distribution corresponding to days with no rain. A simulation study reported in Section 5 allows us to check that the stochastic expectation maximization algorithm proposed for inference works properly on artificial data. Section 6 presents results for temperature and precipitation forecasts based on combining the European, Canadian and US ensembles for five watersheds in Quebec, Canada. The results indicate that post-processed ensembles are much better calibrated and generally sharper than the raw ensembles for the watersheds under study. Section 7 provides a summary and discusses perspectives of future research.

2 Exchangeability

Suppose that the labeling of the ensemble members has no impact on our prior beliefs: their joint distribution will remain invariant to relabeling the members. Such an hypothesis of exchangeability seems plausible: at least this would ideally represent the desiderata for the proficient meteorologist willing to tune his earth model initial-condition perturbations such that the ensemble members do not exhibit systematically persisting figures over time. Fraley et al. [2010] rely on exchangeability to assume in Eq (1) $\pi_k = \frac{1}{K}$ and $\boldsymbol{\theta}_k = \theta \quad \forall k \in \{1, \ldots, K\}$, and so do other authors with respect their own favorite post-processing technique. But the concept of exchangeability should be farther exploited since it provides formal means to construct operational statistical models of ensemble based strictly on judgments of invariance under the relabeling of the members.

Consider a K-sample $(X_1, X_2, \ldots, X_k, \ldots, X_K)$, such that there exists a random variable Z (with pdf $g(.)$) allowing to write the joint distribution of the X_k, $k \in \{1, \ldots, K\}$ as a mixed effect model:

$$f(x_{1:K}) = \int_z g(z) \left(\prod_{k=1}^{K} f(X_k = x_k | Z = z) \right) dz. \tag{3}$$

Given the random effects Z, the X_k, $k \in \{1, \ldots, K\}$, are independent and identically distributed according to $f(x|Z)$. The joint distribution of the variables $(X_1, X_2, \ldots, X_k, \ldots, X_K)$, remains invariant under a relabeling permutation of the components of the mixture: they are exchangeable. Bruno de Finetti's representation theorem (de Finetti [1931, 1937]) and the work from his followers [Hewitt and Savage, 1955] prove the difficult reciprocal: under technical conditions of regularity valid for a theoretically infinite sequence of exchangeable members, exchangeability means conditional independence and yields to Eq (3). Note that exchangeability does permit marginal dependence between members; for example in the Gaussian case, members must have the same mean and the same variance, but they can be correlated with one another, as long as all correlations are equal (and positive). Exchangeability is especially important for modeling ensemble members by its realism as well as its parsimony. Moreover, there exists a very strong *a priori* argument in favor of a structured model of the ensemble members (the $X'_k s$ in eq 1) around a latent single conditioning variable that would explain their dependencies. It is the very objective of the simulation method of the hydrometeorological model of the earth system to target, through a given ensemble of exchangeable simulations, the estimation of a latent meaningful "physical" variable for all members. Such a Z (in Eq (3)) can be interpreted as a statistical synthesis of this latent variable common to the members of the ensemble. This interpretation is furthermore often confined by the strength of the first component obtained through a preliminary PCA (Principal Component Analysis) of the ensemble.

One may object that some physical processes to generate ensembles do not produce exchangeable members: for instance, even and odd numbered members from the Canadian ensemble (CMC-EPS) for precipitations are obtained by different meteorological models[3] (see second panel of Fig 7) and would not pass the tests for accepting exchangeability for the whole CMC-EPS.

The rank statistics lead to a possible exchangeability test for ensembles. A necessary condition for exchangeability is indeed that each member of the ensemble occupies all possible ranks (i.e. $\{1, \cdots, K\}$) in a roughly equal way. We therefore check that the frequency of occupancy of a rank does not deviate too far from $\frac{1}{K}$ for each member (with a χ^2 test).

In the case exchangeability is not an acceptable assumption for the whole ensemble, one can subset the ensemble into exchangeable parts to be dealt with as new ensembles. We thus get back to the case of exchangeability of members belonging to different ensembles. This will not prevent using the models hereafter since, in this paper, we develop a method allowing to incorporate the information conveyed by multiple ensembles into the statistical analysis.

3 Model and inference in the Gaussian case

As in most post-processing techniques, we consider univariate models to obtain a calibrated marginal distribution for each site, each meteorological variable (here, the temperature), and for each lead time. Indeed, the ensemble is generally assumed to be a sufficient summary statistics as far as prediction is concerned. The spatial, temporal and inter-variable dependencies are recovered by using empirical copulas.

[3] `http://collaboration.cmc.ec.gc.ca/cmc/ensemble/doc/info_geps_e.pdf"`

In this section, we consider the simple situation in which the variable to be predicted, Y_t as well as the corresponding ensemble of predictors, $\mathbf{X}_t$ (produced a time $t-h$, h being the considered lead-time), can be assumed jointly (and conditionally) Gaussian. This is notably the case for daily temperature observations over a season but it may also concern other continuous variables after a suitable normalizing transformation.

3.1 Multi-ensemble Exchangeable Gamma Normal Model

Let E be the number of forecast sources (e.g. the ensembles from several meteorological centers) and K_e the number of members within ensemble e. When the forecasting system delivers a single forecast, for instance in the case of an expert issuing a deterministic forecast, we simply set $K_e = 1$. From now on, we also make the convention that $e = 0$ will denote the variable to be predicted and we will sometimes conveniently write: $X_{0,t} = Y_t$, with $K_0 = 1$, as if it were a peculiar ensemble with a single member. This notational trick will be useful for the inference part, where it makes sense since past observations of the target provide information about the unknown of our model just as ensemble members do. Of course, when forecasting, predictand Y_t and predictors $X_{e,k,t}$ will keep their non symmetrical roles. We propose the following model for a given lead time and a given location: for any e, k, t,

$$\begin{cases} (X_{e,k,t}|Z_t) & = a_e + b_e Z_t + c_e \varepsilon_{e,k,t} \\ (Y_t|Z_t) = (X_{0,t}|Z_t) & = a_0 + Z_t + \varepsilon_{0,1,t} \\ \left(\varepsilon_{e,k,t}|\omega_t^{-2}\right) & \overset{ind}{\sim} \mathcal{N}\left(0, \omega_t^2\right) \\ \left(Z_t|\omega_t^{-2}\right) & \overset{ind}{\sim} \mathcal{N}\left(0, \lambda\omega_t^2\right) \\ \omega_t^{-2} & \overset{iid}{\sim} \Gamma\left(\alpha, \beta\right) \end{cases}, \tag{4}$$

where $X_{e,k,t}$ denote the k^{th} member of the ensemble e at time t, Z_t and ω_t^2 are the corresponding latent variables (forming the bedrock of the exchangeability property) upon which the ensemble members of a given ensemble e are conditionally independent. These latent variables Z_t and ω_t^2 are assumed to be independent across time. $\Gamma(\alpha, \beta)$ is the gamma distribution with parameters α and β where α and β, as well as λ and $\{a_e, b_e, c_e\}_{e\in\{0,\dots,E\}}$, are parameters to be estimated. These parameters are then specific to the considered lead-time and location. Identifiability constraints impose $b_0 = c_0 = 1$. Figure 2 shows the corresponding directed acyclic graph (DAG) for the model.

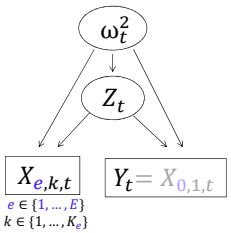


Figure 2: Directed acyclic graph of the model given by Eq 4. $X_{e,k,t}$ is the k^{th} member of the ensemble e at time t and K_e is the number of members from the ensemble e, $Y_t = X_{0,K_0=1,t} = X_{0,t}$ relates to the targeted variable to forecast, and Z_t and ω_t^2 are the model backbone latent variables. Times are assumed to be independent.

The first latent variable, Z_t, can be interpreted as some *hidden global state of the atmosphere*, as *seen* by a meteorological simulator issuing a forecasting ensemble. We make the additional assumption that this pivotal quantity is the same for all ensembles $e = 1, \dots, E$. It makes sense to think that dispersed members yield a large uncertainty on the latent variable Z_t. Since this dispersion is not constant over time, the second latent variable, ω_t^2, is useful in the model to account for this variation. This is related to using of the variance term in the EMOS model [Gneiting et al., 2005]. A meteorological interpretation of this second latent variable ω_t^2 would be something like the underlaying *turbulent atmospheric condition* (as *encoded* by all meteorological simulators). The *adhoc* dependence between Z_t and ω_t^2 as specified by Eq 4 greatly facilitates inference (through the use of Gamma-Normal conjugacy) and therefore leads to a fast algorithm, which is useful in an operational context, where inference can be conducted within a moving window. The meaning of parameters a, b, c, α, β from the model given by Eq 4 is straightforward:

- The difference $a_e - a_0$, $e > 0$ gives the additive bias for the forecasting ensemble e, to be compared to 0.
- The ratio $\frac{b_e}{b_0}$, $e > 0$ is the multiplicative bias of the forecasting ensemble e. Since $b_0 = 1$ for identifiability, the value b_e is directly to be compared to 1. Additive and multiplicative biases may partly compensate one another.
- For parameter c, the ratio $\frac{c_e}{c_0} = c_e$, $e > 0$ (parameter c_0 being fixed to 1) will be understood as a dispersion bias for the predictors. A ratio greater than 1 can be interpreted as an over-dispersion of the predicting ensemble e .
- The ratio $\frac{\beta}{\alpha - 1}$ corresponds to the expected value of ω_t^2 which rules how far the quantity to forecast Y_t can occur from the latent variable Z_t. It is therefore expected that this ratio will increase with the lead time of the forecast, because ensembles generally become less and less informative when the forecasting horizon grows.

The model given in this section fulfills parsimony and integration of multiple sources of information: each additional ensemble only needs three parameters to be included within the multi-ensemble gamma Normal exchangeable model given by Eq (4).

3.2 Inference

In what follows, the parameters to be estimated are denoted by $\boldsymbol{\theta} = (\alpha, \beta, \lambda, \mathbf{a}, \mathbf{b}, \mathbf{c})$ where $\mathbf{a} = (a_e)_{e \in \{0,\dots,E\}}$, $\mathbf{b} = (b_e)_{e \in \{1,\dots,E\}}$ and $\mathbf{c} = (c_e)_{e \in \{1,\dots,E\}}$, recalling that $b_0 = c_0 = 1$.

We moreover use Gelfand's bracket notations for probability distributions [Gelfand and Smith, 1990] and we denote by $(\mathbb{X}, \mathbf{Y})$ the set of predictors $\mathbf{X}_t$ and observations Y_t acquired over time during the learning period and $\mathbf{Z}$ and $\boldsymbol{\omega}^2$ the sets of latent variables.

Assuming that the parameters remain the same over a learning period close to or homogeneous to the prediction period, the EM algorithm [Dempster et al., 1977] is an effective instrument for estimating the parameters of this multivariate normal model with random effects. The E-step is tractable since, for any t, the conditional distribution $\left[Z_t, \omega_t^{-2} | \mathbf{X}_t, Y_t; \boldsymbol{\theta}\right]$ follows a normal-gamma distribution. In the M-step, some parameter updatings have explicit formulas and another relies on a numerical optimization procedure. The proof and explicit formulas for updating the parameters are provided in Appendix A. We denote by $\boldsymbol{\theta}^{(h)} = \left(\alpha^{(h)}, \beta^{(h)}, \lambda^{(h)}, \mathbf{a}^{(h)}, \mathbf{b}^{(h)}, \mathbf{c}^{(h)}\right)$ the current value of the parameters.

Algorithm 1: EM algorithm for estimating parameters in Model (4)

Initialization: Set $\boldsymbol{\theta}^{(0)}$ by a method of moments.

repeat

E-step Compute needed moments of latent variables in $\mathbf{Z}$ and $\boldsymbol{\omega}^2$ with respect to the current parameter values ($\boldsymbol{\theta}^{(h)}$).

M-step Update the current parameter values:

$$\boldsymbol{\theta}^{(h+1)} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{[\mathbf{Z}, \boldsymbol{\omega}^{-2} | \mathbb{X}, \mathbf{Y}; \boldsymbol{\theta}^{(h)}]} \left(\log \left(\left[\mathbb{X}, \mathbf{Y}, \mathbf{Z}, \boldsymbol{\omega}^{-2}; \boldsymbol{\theta} \right] \right) \right)$$

until $\left\| \boldsymbol{\theta}^{(h+1)} - \boldsymbol{\theta}^{(h)} \right\| < \varepsilon$

3.3 Forecasting

For a new time t' with predictors $\mathbf{X}_{t'}$, the forecast for $Y_{t'}$ is provided by the predictive distribution of $(Y_{t'} | \mathbf{X}_{t'})$ (the forecasting target in the operational systems considered here), which is given in the next proposition.

Proposition 1. *Under Model 4, for a new time t', the predictive distribution $(Y_{t'} | \mathbf{X}_{t'} = \mathbf{x}_{t'})$ follows a Student distribution with scale parameter $\sqrt{\frac{(\lambda'' + 1)\beta''_{t'}}{\alpha''}}$, location parameter $a_0 + m''_{t'}$ and $2\alpha''$ degrees of*

freedom where

$$\begin{array}{rcl} \alpha'' & = & \alpha + \frac{\sum_{e=1}^{E} K_e}{2}, \\ \lambda''^{-1} & = & \sum_{e=1}^{E} K_e b_e^2 c_e^{-2} + \lambda^{-1}, \\ m''_{t'} & = & \lambda'' \cdot \sum_{e=1}^{E} c_e^{-2} b_e K_e \left(\bar{x}_{e,t'} - a_e\right), \\ \beta''_{t'} & = & \beta + \frac{1}{2}\left\{\sum_{e=1}^{E}\sum_{k=1}^{K_e} c_e^{-2}(x_{e,k,t'} - a_e)^2 - m''^2_{t'}\lambda''^{-1}\right\}. \end{array}$$

Proof. By using the conjugacy properties of the normal gamma model as in the E-step of Algorithm 1, we obtain that $\left(Z_{t'}|\omega_{t'}^{-2}, \mathbf{X}_{t'}\right)$ follows a normal distribution $\mathcal{N}(m''_{t'}, \lambda''\omega^2_{t'})$ and $\left(\omega_{t'}^{-2}|\mathbf{X}_{t'}\right)$ follows a gamma distribution $\Gamma\left(\alpha'', \beta''_{t'}\right)$ where parameters are determined by identification. Moreover, we have from Model (4):

$$\begin{array}{rcl} \left(Y_{t'}|Z_{t'}\right) & = & a_0 + Z_{t'} + \varepsilon_{0,1,t'} \\ \left(\varepsilon_{0,1,t'}|\omega^2_{t'}\right) & \sim & \mathcal{N}\left(0, \omega^2_{t'}\right). \end{array}$$

Then, $\left(\frac{Y_{t'}}{\sqrt{\lambda''+1}}|\omega^2_{t'}, \mathbf{X}_{t'}\right) \sim \mathcal{N}\left(\frac{a_0+m''_{t'}}{\sqrt{\lambda''+1}}, \omega^2_{t'}\right)$ and by using the distribution of $\left(\omega_{t'}^{-2}|\mathbf{X}_{t'}\right)$ we obtain the announced result. □

Remark 1. *The predictive mean and the predictive variance are given respectively by:*

$$\begin{array}{rcl} \mathbb{E}\left(Y_{t'}|\mathbf{X}_{t'} = \mathbf{x}_{t'}\right) & = & a_0 + \lambda'' \sum_{e=1}^{E} \frac{b_e}{c_e^2} K_e \left(\bar{x}_{e,t'} - a_e\right), \\ \mathbb{V}\left(Y_{t'}|\mathbf{X}_{t'} = \mathbf{x}_{t'}\right) & \propto & \beta + \frac{1}{2}\left\{\sum_{e=1}^{E} c_e^{-2}\sum_{k=1}^{K_e}(x_{e,k,t'} - a_e)^2 - \lambda''\left\{\sum_{e=1}^{E}\frac{b_e}{c_e^2}K_e\left(\bar{x}_{e,t'} - a_e\right)\right\}^2\right\}. \end{array} \quad (5)$$

This predictive distribution is very similar to the EMOS one, (see Eq (2)), for which the predictive expectation is expressed linearly as a function of the mean of the members and the predictive variance as a function of the ensemble variance. In Eq (5), it appears that a member from ensemble e has an even greater impact on the forecast as the $\frac{b_e}{c_e^2}$ ratio is large. Therefore, we define the *contribution* of a member of the ensemble e to the final forecast by:

$$contrib_e = \frac{\frac{b_e}{c_e^2}}{\sum_{e'=1}^{e'=E} K_{e'}\frac{b_{e'}}{c_{e'}^2}}.$$

4 An extension of the multi-ensemble exchangeable Gamma Normal model to the precipitation case

In this section, we investigate an adaptation of the post-processing method based on the exchangeability hypothesis for precipitation-like variables. These variables cannot be assumed to be normally distributed: they exhibit a mixed nature with a discrete component at zero and a positive continuous component. In the long term, we wish to be able to jointly post-process temperature variables and rainfall type variables, this is one of the reasons why we seek to remain within the convenient framework of the Gaussian family. The approach proposed herein can be viewed as a Tobit regression (Tobin, 1958 ; Chib, 1992) and is similar to techniques presented in Scheuerer and Hamill [2015] and Thorarinsdottir and Gneiting [2010].

4.1 Multi-ensemble Multilevel Exchangeable Tobit model

The underlying idea of the following model is that precipitation (as observations with a zero discrete component and a continuous positive component), would be some left censorship from a continuous (latent) variable [Allard, 2012]. Based on this idea, some work has been already undertaken to develop a post-processing method for precipitation forecasts by [Schultz et al., 2010]. In this work, the latent continuous variable associated with precipitation has a physical meteorological interpretation and is called *pseudo-precipitation*. We leave aside any physical interpretation and assume that these pseudo-precipitations are Gaussian after an appropriate invertible transformation $f_{\mathcal{N}}$ such as the Box-Cox one [Box and Cox, 1964]. The model we propose for such normalized pseudo-precipitation is the same as that proposed for temperature variables.

Let Y'_t and $\mathbf{X}'_t$ be the precipitation to forecast and its predictors. Y_t and $\mathbf{X}_t$ become in this section latent variables: they correspond to the underlaying normal pseudo-precipitation and its normal pseudo-ensembles of predictors. The model is as follows for a given lead time and a given location: for any e, k, t:

$$\begin{cases} (X_{e,k,t}|Z_t) & = a_e + b_e Z_t + c_e \varepsilon_{e,k,t} \\ (Y_t|Z_t) = (X_{0,1,t}|Z_t) & = a_0 + Z_t + \varepsilon_{0,1,t} \\ \left(X'_{e,k,t}|X_{e,k,t}\right) & = \mathbb{I}_{X_{e,k,t}>\nu} f_{\mathcal{N}}^{-1}\left(X_{e,k,t}\right) \\ \left(\varepsilon_{e,k,t}|\omega_t^{-2}\right) & \overset{ind}{\sim} \mathcal{N}(0, \omega_t^2) \\ \left(Z_t|\omega_t^{-2}\right) & \overset{ind}{\sim} \mathcal{N}(0, \lambda\omega_t^2) \\ \omega_t^{-2} & \overset{iid}{\sim} \Gamma\left(\alpha, \beta\right) \end{cases} \tag{6}$$

$\mathbb{I}$ denotes the indicator function. Z_t and ω_t^2 are the latent backbone of this multilevel exchangeable model and are assumed to be independent across time. Conditionally upon (Z_t, ω_t^2), the ensemble members are iid within each ensemble. α, β, λ, ν and $\{a_e, b_e, c_e\}_{e\in\{0,\dots,E\}}$ are parameters to be estimated (specifically to each lead time and location). Again for identifiability concerns, $b_0 = c_0 = 1$. The random variables $(X_{e,k})_{e\in 0,\dots,E\ k\in 1,\dots,K_e}$ are now latent (yet observed when greater than ν). Figure 3 shows the DAG (Directed Acyclic Graph) that corresponds to the precipitation model.

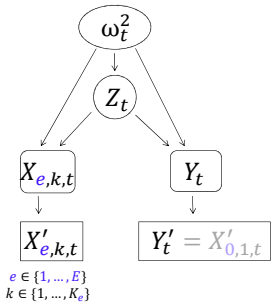


Figure 3: Direct acyclic graph of the post-processing model for precipitation. $X'_{e,k,t}$ denotes the k^{th} member of ensemble e at time t, K_e is the size of the ensemble e, $Y'_t = X'_{0,K_0=1,t} = X'_{0,t}$ denotes the variable to forecast and $\mathbf{X_t}$, Y_t, Z_t and ω_t^2 are latent variables (partially observed in the case of $\mathbf{X_t}$ and Y_t). Times are assumed to be independent.

4.2 Inference

As for the model of the previous section, the EM algorithm is tailored for estimating the parameters of our exchangeable Tobit model. Beforehand, we proceed with the estimation of the normalizing transformation from historical precipitation data.

Normalisation parameters

The selection of an appropriate normalizing transformation $f_{\mathcal{N}}$ is an important issue. In this work, we consider the power transform: $f_{\mathcal{N}}(x') = x'^{\gamma}$, where γ is a parameter to be estimated. We choose the same transformation parameter regardless of the precipitation ensemble considered, the one estimated from the observed precipitation values (variable to predict). Thus, in our model, $f_{\mathcal{N}}$ does not depend on the ensemble e. Consequently, we can use all historical precipitation data available for inferring parameter γ. Obviously, for precipitations, it is essential to ensure that the inverse transformation, $f_{\mathcal{N}}^{-1}$, takes positive values on $]\nu, +\infty[$. In this work, we set $\nu = 0$ which corresponds to a very refined sensitivity of the rain gauge, with non-zero values of observed rainfall very close to 0. The goal is therefore to find a value for γ such that Y'^{γ} can be assumed to follow a normal distribution and is left censored at 0.

Assuming such a model with temporal independence of the precipitation phenomenon, the log-

likelihood, can be written as follows:

$$\begin{aligned}\mathcal{L}\left(\{y'_t\}_t\right) &= \sum_{t \text{ s. t. } y'_t>0} \left(\log\left[Y'_t = y'_t\right]\right) + \#\{t, y'_t = 0\}\log([Y'_t = 0]) \\ &= \sum_{t \text{ s. t. } y'_t>0} \left(\log\left[Y'^{\gamma}_t = y'^{\gamma}_t\right] + (\gamma-1)\log\left(y'_t\right) + \log\left(\gamma\right)\right) + \#\{t, y'_t = 0\}\log([Y_t \leq 0]) \\ &= \sum_{t \text{ s. t. } y'_t>0} \left(\log\psi\left(y'^{\gamma}_t; \mu, \sigma^2\right) + (\gamma-1)\log\left(y'_t\right) + \log\left(\gamma\right)\right) + \#\{t, y'_t = 0\}\log(\Psi(0; \mu, \sigma^2)),\end{aligned}$$

where μ et σ are also parameters to be estimated, $\psi\left(x; \mu, \sigma^2\right)$ and $\Psi\left(x; \mu, \sigma^2\right)$ are respectively the pdf and the cumulative distribution function at x of a Gaussian distribution with mean μ and variance σ^2 and # denotes the cardinal of a set. In practice, parameters μ, σ^2 and γ are obtained by maximizing the likelihood with a numerical optimization method (the Nelder-Mead procedure implemented in `R`). The power transformation is then applied to the observations and the corresponding ensemble forecasts.

Other parameters: the Stochastic EM algorithm

The E-step of the EM algorithm requires to compute, for any t, the conditional distribution function of $(\mathbf{X}_t, Y_t, Z_t, \omega_t^{-2}|\mathbf{X}'_t, Y'_t)$, which is not tractable. The distribution of $\left(Z_t, \omega_t^{-2}|\mathbf{X}_t, Y_t, \mathbf{X}'_t, Y'_t\right)$ is the same as the distribution of $\left(Z_t, \omega_t^{-2}|\mathbf{X}_t, Y_t\right)$ and the distribution of $\left(\mathbf{X}_t, Y_t|Z_t, \omega_t^{-2}, \mathbf{X}'_t,\ Y'_t\right)$ is given by: for any e, k, t (included $e = 0$ that is $X_{0,1,t} = Y_t$),

$$\left[X_{e,k,t}|Z_t, \omega_t^{-2}, X'_{e,k,t}\right] = \begin{cases} \mathbb{I}_{\left\{X_{e,k,t}=f_{\mathcal{N}}(X'_{e,k,t})\right\}} \text{ if } X'_{e,k,t} > 0 \\ \psi_{<\nu}\left(X_{e,k,t}; a_e + b_e Z_t, c_e^2\omega_t^2\right) \text{ if } X'_{e,k,t} = 0 \end{cases} \tag{7}$$

where $\psi_{<\nu}\left(x; \mu, \sigma^2\right)$ denotes the Gaussian pdf with mean μ and variance σ^2 truncated to the right at ν. Therefore, we can add a simulation step (S-step) before the E- step in the inference algorithm. This leads to a partially stochastic EM algorithm [Broniatowski et al., 1983, Celeux and Diebolt, 1985]. The SEM algorithm and the Gibbs algorithm used in S-step are provided hereafter.

Algorithm 2: SEM algorithm for estimating parameters in Model (6)

Initialization: Set $\boldsymbol{\theta}^{(0)}$ by a method of moments.
repeat

S-step For each time t of the learning set, Simulate $(\mathbf{X}_t^S, Y_t^S)$ with respect to the conditional distribution $\left[\mathbf{X}_t, Y_t|\mathbf{X}'_t, Y'_t; \boldsymbol{\theta}^{(h)}\right]$ by Algorithm 3.

E-step Compute needed moments of latent variables in $\mathbf{Z}$ and $\boldsymbol{\omega}^2$ with respect to the current parameter values $(\boldsymbol{\theta}^{(h)})$ and $(\mathbb{X}^S, \mathbf{Y}^S)$.

M-step Update the current parameter values:

$$\boldsymbol{\theta}^{(h+1)} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\left[\mathbf{Z}, \boldsymbol{\omega}^{-2}|\mathbb{X}^S, \mathbf{Y}^S; \boldsymbol{\theta}^{(h)}\right]}\left(\log\left(\left[\mathbb{X}^S, \mathbf{Y}^S, \mathbf{Z}, \boldsymbol{\omega}^{-2}; \boldsymbol{\theta}\right]\right)\right)$$

until $\left\|\boldsymbol{\theta}^{(h+1)} - \boldsymbol{\theta}^{(h)}\right\| < \varepsilon$

Algorithm 3: Gibbs algorithm for simulating latent variables $(\mathbf{X}, Y)$ conditionally to $(\mathbf{X}', Y')$

For a current value of the parameters $\boldsymbol{\theta}^* = (\alpha^*, \beta^*, \lambda^*, \mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*)$.
Initialization: Simulate $\omega^{-2,(1)}$ from $\Gamma(\alpha^*, \beta^*)$ then $Z^{(1)}$ from $\mathcal{N}(0, \lambda^*\omega^{2,(1)})$.
for *i in 1 to* N_{iter} **do**

1. For each (e, k) such that $X'_{e,k} = 0$, simulate $X_{e,k}^{(i)}$ from $\left[X_{e,k}|Z^{(i)}, \omega^{-2,(i)}, X'_{e,k}; \boldsymbol{\theta}^*\right]$.
2. Simulate $(Z^{(i+1)}, \omega^{-2,(i+1)})$ from $\left[Z, \omega^{-2}|\mathbf{X}^{(i)}, Y^{(i)}; \boldsymbol{\theta}^*\right]$.

Return: $(\mathbf{X}^S, Y^S) = (\mathbf{X}^{(N_{iter})}, Y^{(N_{iter})})$.
end

Note that in Algorithm 3, for each t, the values of $X_{e,k,t}$ corresponding to $X'_{e,k,t} \neq 0$ are set as $X_{e,k,t} = f_{\mathcal{N}}(X'_{e,k,t})$. The number of iterations N_{iter} has to be chosen large enough to ensure that the Gibbs algorithm has converged.

4.3 Forecasting

In order to simulate $Y'_{t'}$ from the conditional distribution of $(Y'_{t'}|\mathbf{X}'_{t'})$, a sample from the distribution of $\left(Z_{t'}, \omega_{t'}^{-2}|\mathbf{X}'_{t'}\right)$ is first drawn. Then, for each couple $\left(Z_{t'}, \omega_{t'}^{-2}\right)^{(i)}$ of this sample, we simulate $Y_{t'}^{(i)}$ from the distribution of $(Y_{t'}|Z_{t'}^{(i)}, \omega_{t'}^{-2,(i)})$ and apply $Y_{t'}^{\prime(i)} = f_{\mathcal{N}}^{-1}\left(Y_{t'}^{(i)}\right)\mathbb{I}_{Y_{t'}^{(i)}>\nu}$. The sample of the distribution of $\left(Z_{t'}, \omega_{t'}^{-2}|\mathbf{X}'_{t'}\right)$ is simulated by using Algorithm 3, where $Y'_{t'}$ is not considered in the conditional distributions in Steps 1 and 2. In this sample, we removed the first iterations of the Gibbs Algorithms which correspond to a burn-in period.

5 Simulation study

In this section we check that, in a realistic framework, i. e. for parameters close to those that would be learned from real data sets, we are able to correctly estimate them with the SEM algorithm described above. Since the γ parameter of the normalizing transformation has been learned separately, we work directly in the normalized space. For this algorithmic experiment, we simulate artificial data according to the Tobit model described in Section 4.1 with the following parameters:

$$\begin{aligned} K &= (1, 10, 35, 1) \\ a &= (a_0, a_1, a_2, a_3) = (0, 1, 0.7, -0.1) \\ b &= (b_0, b_1, b_2, b_3) = (1, 1.1, 1, 0.9) \\ c &= (c_0, c_1, c_2, c_3) = (1, 0.8, 0.7, 1.1) \\ \alpha &= 2.5, \quad \beta = 3, \quad \lambda = 0.5, \end{aligned}$$

where the first element of each vector is relative to the observations (referred to with index 0). We generate 100 artificial datasets of 200 elements each. The SEM algorithm described above is then run on the 100 first elements of each of the 100 datasets for inference with the initial value of the parameters chosen by the method of moments. Note that parameters b_0 and c_0 are not estimated: they are set to 1. The SEM algorithm is launched for 1000 steps and, within each iteration, the Gibbs algorithm carries out 4 iterations, which appears to be sufficient in this case given the good estimation results.

The distributions of parameter estimates obtained from the 100 simulated datasets are illustrated in the form of boxplots on Figure 4. Whatever the parameter considered, the resulting estimate does not show any significant bias.

We then assess the performance of the corresponding forecasts. We run our forecasting method on the 100 last elements of each of the 100 simulated data sets, both with the parameters used for simulations ("oracle" forecasts) and with the estimated parameters ("prediction" forecasts). We compare those forecasts to those obtained by considering the raw ensembles as samples from probabilistic forecasts. Guided by [Gneiting et al., 2007], we apply verification tools such as the rank histogram and the continuous ranked probability score (CRPS) to evaluate the performance of those forecasts. The rank histogram assesses their reliability, while the CRPS evaluates both their reliability and their sharpness. A rank histogram is computed from the forecasts on each of the 100 datasets. All these ranks histograms are summarized by the p-values obtained by comparing them to flat histograms through multinomial goodness-of-fit Chi-squared tests. The results of those tests show a good reliability of our forecasts and an even better reliability of the oracle forecasts, as illustrated in Figure 5 (left-hand-side). The reliability of Ensemble 3 (which is a single member ensemble) is omitted. The CRPS (Figure 5, right-hand-side) shows a better performance of our forecasts compared with the ones obtained from the raw ensembles. Note that in the case of the third ensemble (deterministic forecast), the CRPS reduces to the Mean Absolute Error (MAE) [Hersbach, 2000]. We also check the reliability of the simulated distributions of the latent variables Z and ω^2 by computing their coverage rates. The median coverage rates of the 88% credible intervals are respectively of 0.70 and 0.87 for prediction and oracle methods in the case of Z, and of 0.78 and 0.86 in the case of ω^2.

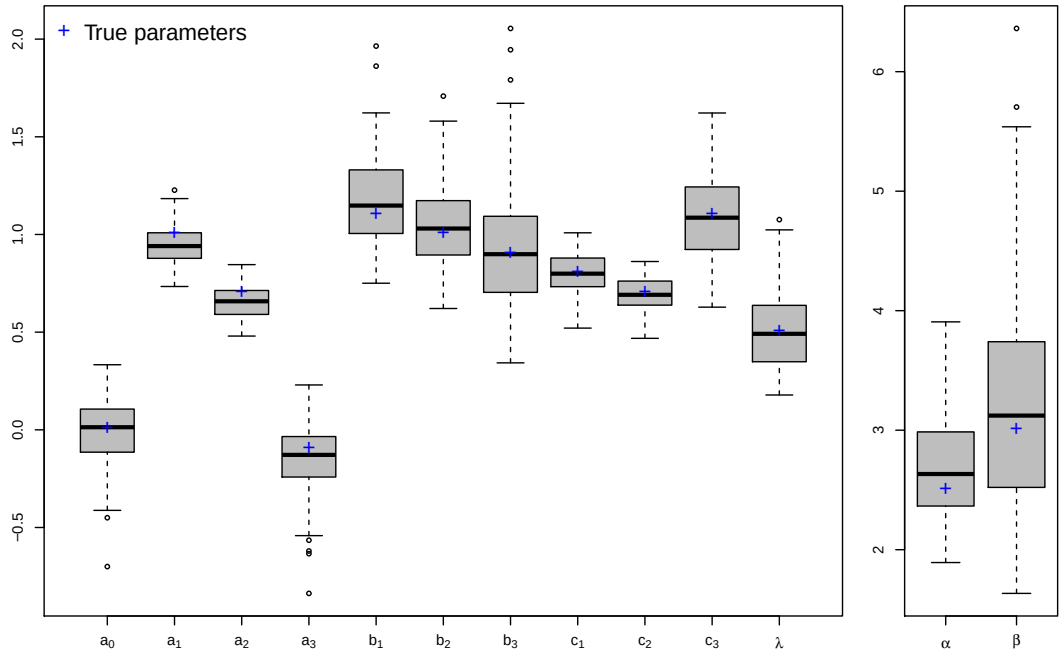


Figure 4: Parameters obtained after inference from 100 simulations of 100 repetitions according to the precipitation model. The parameters used for the simulation are represented by the blue crosses.

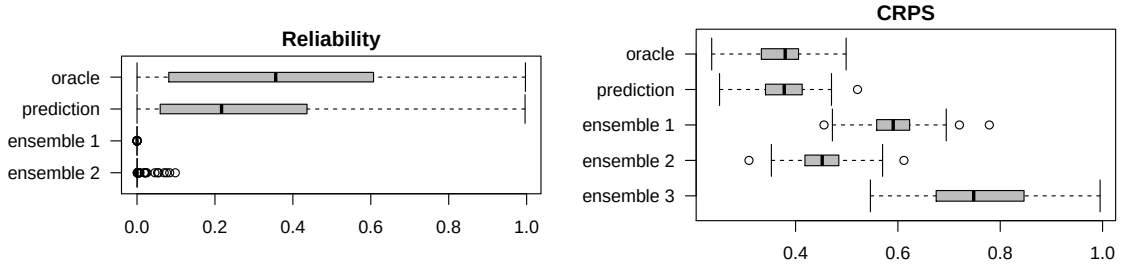


Figure 5: Reliability (left-hand-side) and CRPS (right-hand-side) of forecasts issued by post-processing approaches computed over 100 datasets simulated according to the precipitation model. Reliability is here summarized by the p-values of the multinomial goodness-of-fit Chi-squared tests which compare the rank histograms to a flat histogram. *oracle* and *prediction* refer to our post-processing method with respectively true and estimated parameters. They are compared with the three raw ensembles considered as probabilistic forecasts (Reliability of Ensemble 3 is omitted since it is a single member ensemble).

6 Application to meteorological data from Québec

The case study developed in this section to illustrate the use of our post-processing approaches concerns ensemble temperature and precipitation forecasts available daily over Hydro-Québec Manicouagan watershed, a major hydropower system. As illustrated on Fig 6, the Manicouagan watershed is subdivided into five subcatchments for which meteorological forecasts are used every day to produce streamflow predictions: upstream to downstream, Manic-5, Petit Lac Manicouagan, Toulnoustouc, Manic-3 and Manic-2. The Manicouagan watershed is located in northeast of the province of Québec, Canada. This water resources system consists of two hydropower plants with reservoirs in parallel (Manic-5 and Toulnustouc) and three downstream run-of-river hydropower plants (Manic-3, Manic-2 and Manic-1). The total installed capacity is 6,202 MW, which is about 17 pourcent of Hydro-Québec's total capacity. In operating the system, generation planners face a variety of decisional problems. Two of those, common to every installation, are safety and the respect of environmental laws and regulations. For the two upstream watersheds, with large reservoirs, the other main concerns are those of long term energy planning and optimization, and efficient releases for the operation of the run-of-river plants, given the inflows on those sub-basins. For the three run-of-river plants, the issue is an efficient scheduling, given the inflows

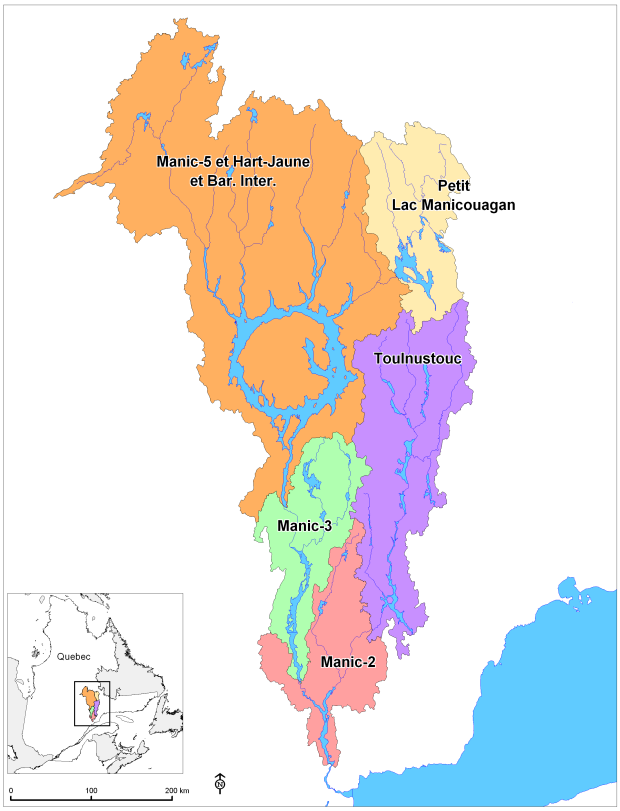


Figure 6: Manicouagan watersheds in Québec, Canada.

on the watersheds and the upstream releases. It is quite clear that a good prediction of the future state of inflows, which highly depends upon weather forecasts, plays a major role on the decisions that will be made, and the efficiency of the operations. The need for reliable ensemble weather forecasts is thus self-evident.

To assess the performance of the post-processing mixed effect models, Hydro-Québec provided us with records of daily meteorological forecasts and corresponding observations for the five watersheds for years 2013 and 2014. The meteorological ensemble forecasts were extracted from the THORPEX Interactive Grand Global Ensemble (TIGGE) database ([Park et al., 2008]). Three daily ensembles for forecasting lead times ranging from 1 to 9 days produced by meteorological global forecast centres were considered:

- the CMC-EPS (Ensemble Prediction System), from the Canadian Meteorological Center (CMC), with 20 ensemble members,
- the NCEP-GEFS (Global Ensemble Forecasting System) from the National Centers for Environmental Prediction (NCEP), with 20 ensemble members,
- the ECMWF-EPS from the European Center for Medium-Range Weather Forecasts (ECMWF), with 50 ensemble members.

Meteorological variables of interest for Hydro-Québec are daily minimal and maximal temperatures, and precipitations. Since the rainfall-runoff model in use is a lumped and conceptual hydrological model [Guay et al., 2018], the raw ensemble forecasts available at grid points belonging to the basins under study, or located adjacent to it, have been averaged to get global watershed values. These are to be compared to the corresponding observed values computed by Hydro-Québec.

In this section, we focus on maximal temperatures to test our Gamma Normal model, and on precipitations, to test our Tobit model. Both models assume that ensemble members are exchangeable within each ensemble. This assumption seems appropriate for the ECMWF-EPS and the NCEP-GEFS, given the way their ensemble members are produced, but it is not for the CMC-EPS. We thus look for

sub-ensembles of members of the CMC-EPS within which exchangeability can be assumed. A rank test shows that, in the case of precipitations, even ensemble members and odd ensemble members constitute two appropriate subgroups (see section 2), we thus treat them as two separate ensembles. In the case of maximal temperatures, however, we could not find any such sub-group. In the absence of a better solution, we consider the whole ensemble.

Figure 7 shows an example of a forecasting situation that has been treated in this case study, a forecast produced on the 30th of April 2014. Raw CMC-EPS, NCEP-GEFS and ECMWF-EPS precipitations and maximal temperature ensemble forecasts for Manic 2 watershed are presented for lead times from 1 up to 9 days. The target values, to be predicted, are indicated by the dotted black lines. On the precipitation example (bottom panel), we decompose the CMC-EPS into its two exchangeable sub-ensembles, odd ensemble members (CMC-EPS-1) and even ensemble members (CMC-EPS-2).

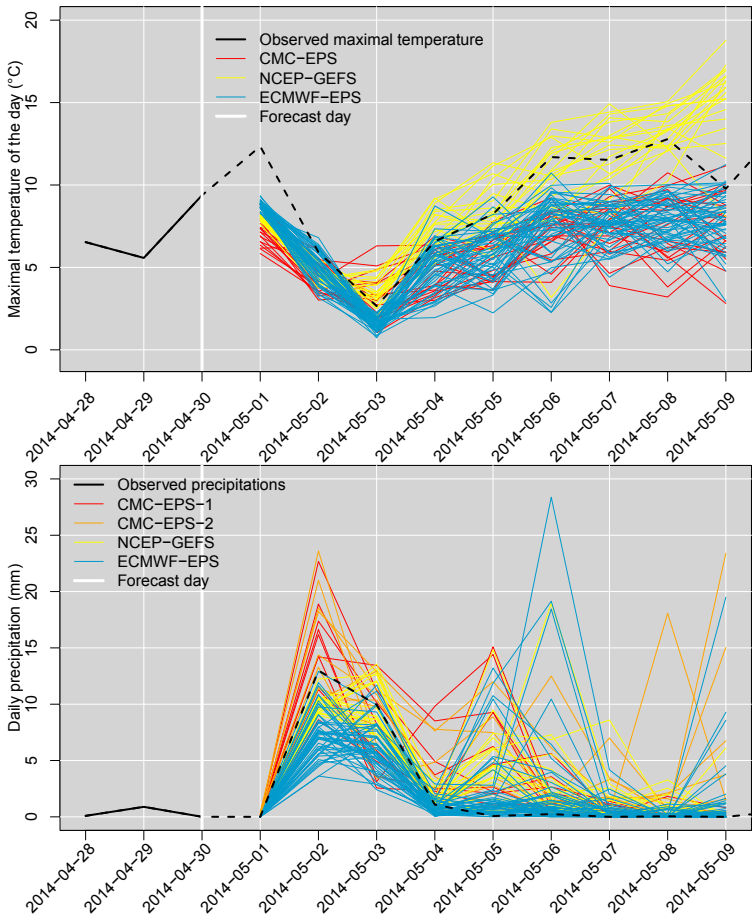


Figure 7: Precipitations and maximal temperature raw ensemble forecasts on the Manic 2 watershed produced the 30 of April 2014 for the nine upcoming days (CMC-EPS, odd members: 1 and even members: 2, NCEP-GEFS and ECMWF-EPS). The target values, to be predicted, are indicated by the dotted black line.

In the following applications, we are seeking to produce reliable predictive distributions for each lead time and each meteorological variables. First, the Gamma Normal post-processing model is applied to daily maximal temperatures ensemble forecasts, and then the Tobit model is applied to precipitations ensemble forecasts. As in Section 5, we apply verification tools such as the rank histogram and the CRPS to evaluate the performance of the post-processing approaches. The CRPS have been calculated from daily predictive distributions produced for 2014, with parameters estimated using observations and raw predictions available for 2013. In both cases, temperatures and precipitations, we first focus on Manic 2 watershed and then extend our results to the four other basins.

6.1 Application to maximal daily temperatures forecasts

Illustration on the Manic 2 watershed

Figure 8 shows the estimated parameters as a function of the forecast lead times. The graph Fig 8(1) illustrates the additive bias of each three ensemble forecasts, given by the difference between parameters a_1, a_2 et a_3, related to the 3 ensembles, and parameter a_0, corresponding to observations. It indicates that the three ensemble prediction systems would have produced negatively biased forecasts during 2013 for the Manic 2 watershed, regardless of the forecasting lead time. The graph Fig 8(2) gives inference results for parameters b that can be interpreted as the forecast multiplicative bias. Its value is less than 1 in the case of CMC-EPS et ECMWF-EPS ensembles, which further amplifies the diagnosis stemming from the first graph. On the other hand, this figure shows a positive multiplicative bias of the NCEP-GEFS ensemble for lead time higher than six days. This could potentially compensate for the negative additive bias previously observed for 2013.

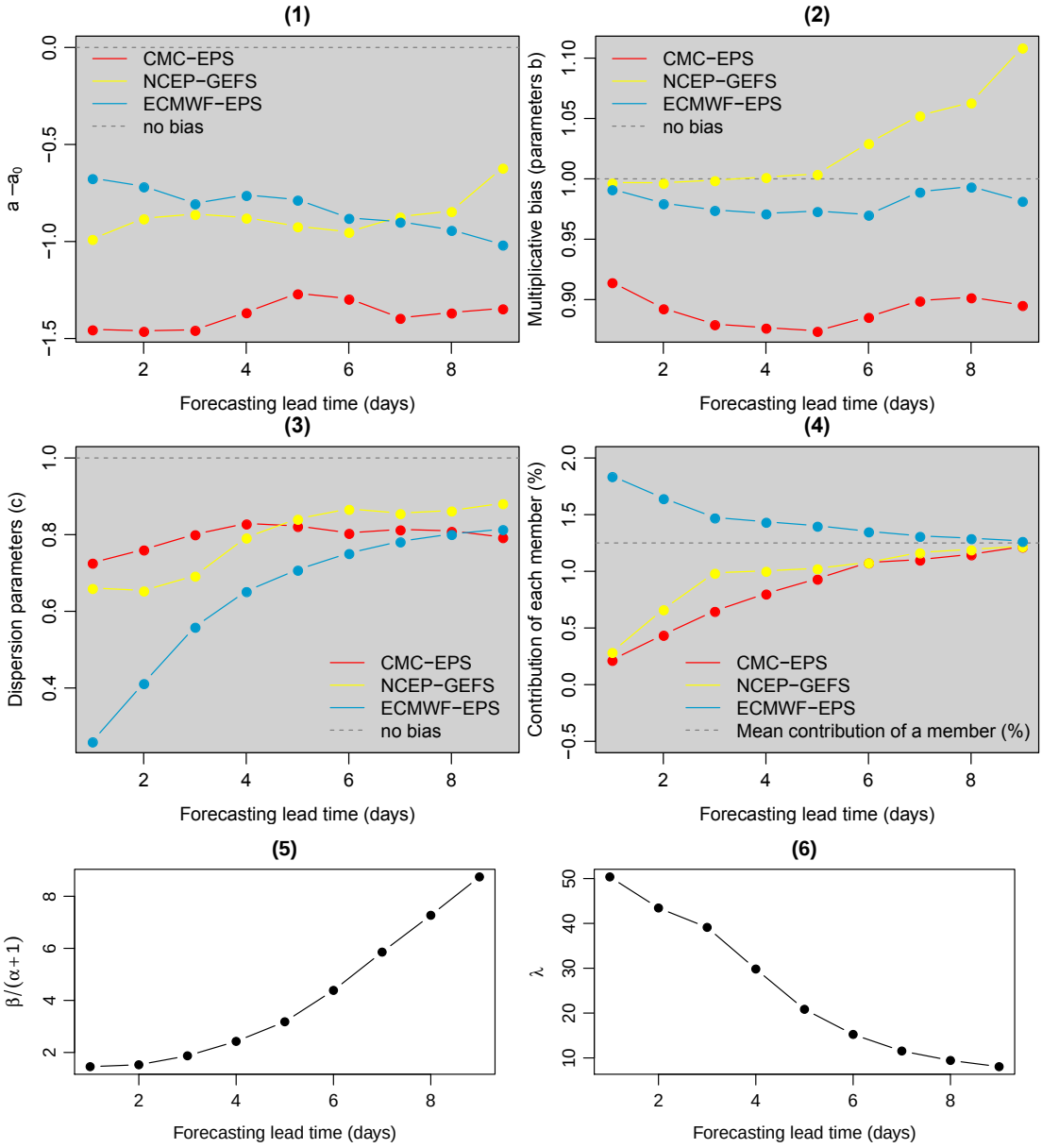


Figure 8: Parameters obtained with the EM algorithm are represented as a function of the forecasting horizon for the Manic 2 watershed with $E = 3$ forecasting sources (CMC-EPS, NCEP-GEFS and ECMWF-EPS). Year 2013 has been used as the learning period.

Fig 8(3) shows the inference results for parameters c which set the inter-member dispersion of the

ensemble relative to the observation one. Note that all values are lower than 1 and increase with the forecast lead time. Therefore, the ensemble produced in 2013 would have been underdispersed, especially for short-term forecasting. As well as the c parameters, the ratio $\frac{\beta}{\alpha-1}$, illustrated on Fig 8(5), increases with the forecast lead time. This ratio corresponds to the expected value of ω_t^2 and therefore settles the variability of the quantity to be forecasted Y_t around the latent variable Z_t. Thus, as expected, the uncertainty blurring Y_t increases with the forecast lead time since forecasts become less informative for longer term prediction. The graph of Fig 8(4) shows the contribution to the final forecast, $contrib_e$, of each member of the ensemble prediction system e. It is expressed in percentage of the total contribution to the overall forecast. We observe that for 2013 at Manic 2 watershed the contribution from the members of the ECMWF-PES is much larger than that from the members of the other ensembles. However, forecasts tend to be similar in terms of contributions when the lead time increases. Recall that the ECMWF ensemble includes 50 members, while the other ones each include 20 members. Therefore most of the information comes from the ECMWF ensemble. This observation is also confirmed by the first bar of Figure 12 where are shown the relative contribution of each forecast source e, given by $K_e \times contrib_e$, averaged over the 9 forecast lead times.

We now evaluate our post-processing approach for daily maximal temperature ensemble forecasts produced in 2014 at the Manic 2 watershed. Figure 9 again shows the forecasting situation of Figure 7, the maximal temperatures forecasts that have been produced on April 30th, 2014.

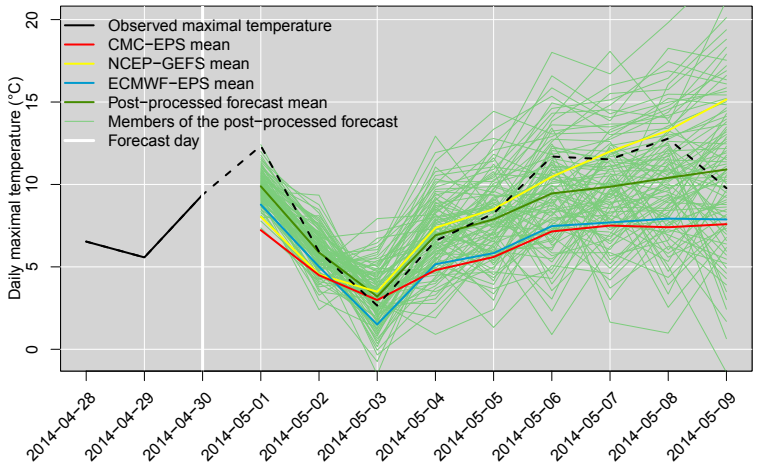


Figure 9: Example of a forecast issued for the daily maximum temperatures of the Manic 2 catchment area with the proposed post-processing method taking NCEP-GEFS, CMC-EPS and ECMWF-EPS as inputs. Predictive scenarios derived from meteorological forecasts by forecast time horizon using the ECC-Q are presented. The maximum daily temperature to be forecast (observed afterward) is indicated by the black dotted line.

The graph shows the averages of the three raw ensembles considered (blue for ECMWF-EPS, yellow for NCEP-GEFS and red for CMC-EPS). The average of the forecasts resulting from post-processing involving these three forecast sources is also presented in green. For this specific forecast situation, the resulting prediction is clearly a compromise between the ECMWF-EPS and CMC-EPS members on the one hand, and NCEP-GEFS on the other. The 90 scenarios of the forecast illustrated by green fine lines are obtained from the post-processed forecasts initially produced independently, lead time by lead time, these outputs being re-ordered using Ensemble Copula Coupling (ECC). This approach consists in copying the rank structure observed in a raw ensemble to produce scenarios from independent points. For instance, if the scenario predicting the highest temperature at time 1 also predicted the lowest temperature at time 2 in the raw ensemble, these two extremes will also be linked in one of the final predictive scenarios. In the ECC-Q version of the ECC, used here, the starting points are quantiles of the marginal predictive distributions [Schefzik et al., 2013].

Based on rank histograms, the 2014 daily post-processed forecasts obtained for the Manic 2 catchment can be considered reliable, regardless of the forecast lead time (figure not shown). The CRPS values presented in Fig 10 support these results. This figure shows the following comparisons :

- the forecasts obtained by considering the raw CMC-EPS ensemble, `CMC-EPS`,

- the forecasts obtained by post-processing the CMC-EPS with our method after learning the model parameters on year 2013 (in this case, $E = 1$), `Post-processed CMC-EPS`,
- the forecasts obtained by post-processing the CMC-EPS with the standard EMOS method, described in Gneiting et al. [2005], `Post-processed CMC-EPS with EMOS`
- the forecasts obtained by considering the members of the CMC-EPS, NCEP-GEFS and ECMWF-EPS ensembles combined in a large ensemble, `Grand Ensemble`,
- the forecasts obtained by post-processing the CMC-EPS, NCEP-GEFS and ECMWF-EPS ensembles, considered as 3 distinct forecast sources ($E = 3$) according to the post-processing method proposed herein whose corresponding parameters are illustrated in Fig 8, `Post-processed Grand Ensemble`.

It is seen that the raw maximal temperature ensembles of 2014 have higher CRPS values for all lead times compared to the corresponding post-processed ones. The raw CMC-EPS forecasts, currently used by Hydro-Québec for inflow forecasting, are significantly improved (according to CRPS) by the proposed post-processing method. One can also observe that its performance is of the same order of magnitude as that of the standard EMOS method, which was expected since these methods give similar forecasts. Furthermore, multi-ensemble post-processing makes it possible to improve the forecasts of the large raw ensemble, in particular for the shorter lead times. Finally, Figure 10 shows, for the Manic-2 catchment, that the combination of several sources of forecasts with statistical post-processing would be the option to be favored. This configuration indeed obtains the smallest CRPS values.

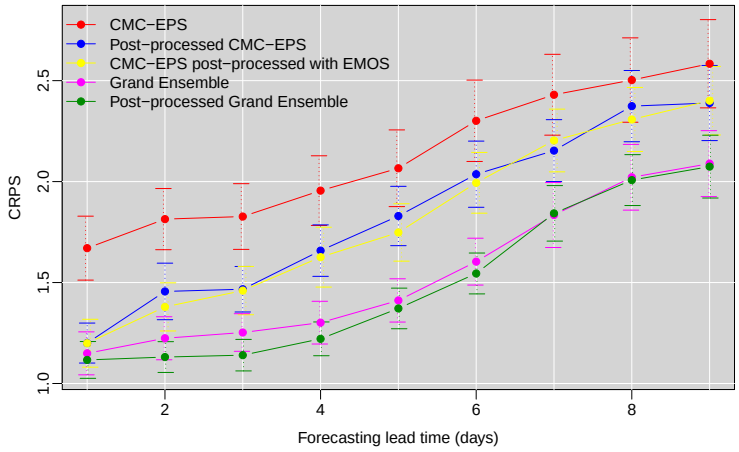


Figure 10: CRPS (in °C) for maximum temperature forecasts at Manic-2 catchment for all 9 lead times with associated bootstrap intervals. The parameters have been estimated using year 2013 and verification has been performed for year 2014.

Extending the results to all watersheds

Based on rank histograms (not shown here), the post-processed maximal temperature daily ensemble forecasts for 2014 can be considered reliable regardless of the catchment and the lead time. According to the CRPS, post-processed ensembles are generally better than raw ensembles. Figure 11 reports the average CRPS values as a function of lead times for each of the five watersheds studied. The main observations drawn from this figure are as follows:

- CRPS values show that post-processing globally improves the reliability and accuracy of forecasts for all watersheds and almost all forecasting lead times. This observation applies to CMC-EPS forecasts (red curve compared to blue curve) as well as general forecasts from several sources (magenta curve compared to green curve). It is therefore in our best interest to post-process ensemble maximum daily temperature forecasts produced for the 5 watersheds of Manicouagan hydropower system.

- According to the CRPS, combining several forecast sources, together with statistical post-processing, reduces the average error of forecast by at least 0.5°C, compared to a single source of raw ensemble. This result can make all the difference when producing hydrological predictions during seasonal transitions (frost in the fall, snowmelt in the spring) since temperature forecasting plays a crucial role at these times of the year.

- Surprisingly, for Toulnustouc watershed, 1 and 2-days ahead forecasts of the post-treated large ensemble obtain CRPS values (green curve) that are greater than those of the corresponding raw ensemble (magenta curve). For the 1 day ahead prediction its performance is even worse than the single source CMC-EPS raw ensemble forecasts (red curve). This deterioration might stem from a problem of non-homogeneity in the dataset: the model parameters fitted on the learning sample may be not appropriate for the validation one. For instance, in Quebec, 2014 was very cold in winter compared to 2013.

- Our approach obtains slightly smaller CRPS values than that of the standard EMOS method for 4 watersheds: Manic 5, Petit Lac Manic, Toulnustouc and Manic 3 (blue curve compared to yellow curve).

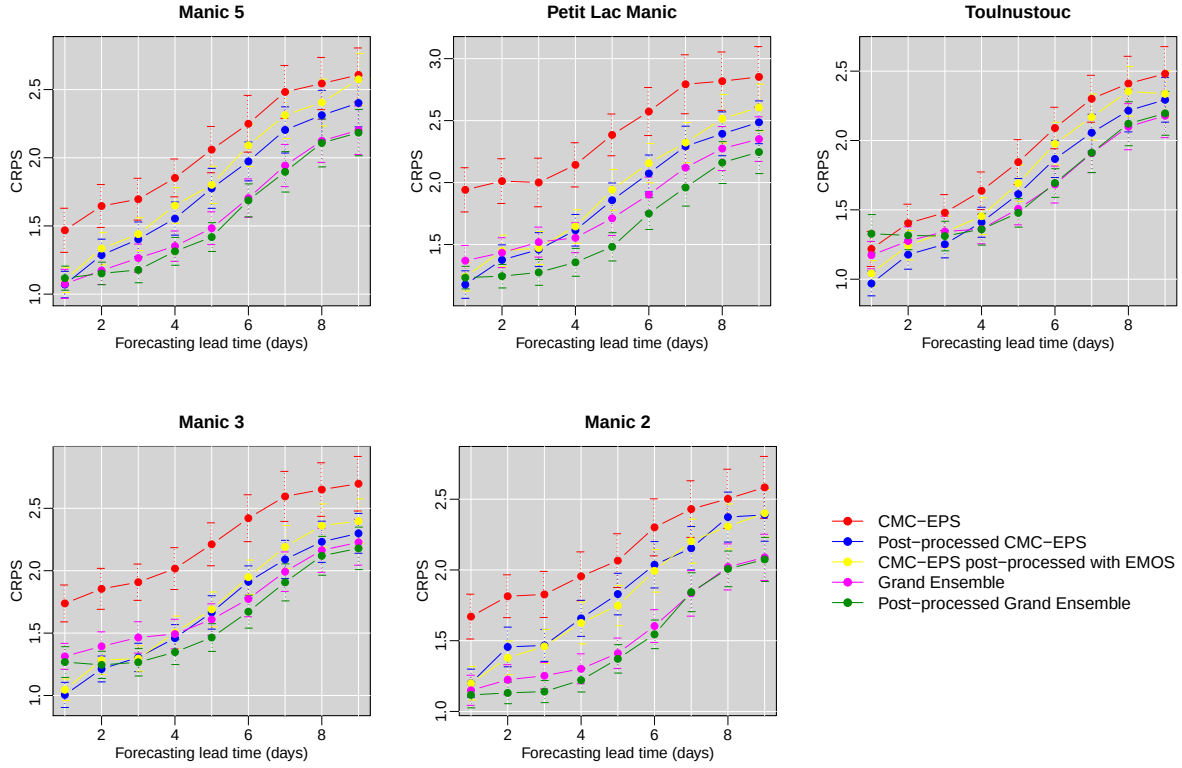


Figure 11: CRPS (in °C) for maximum temperature forecasts for the five catchments for all 9 lead times with associated bootstrap intervals. The parameters have been estimated using year 2013 and verification has been performed for year 2014.

Finally, Figure 12 shows that observations made on the relative contributions of forecast sources for Manic 2 extend to the four other watersheds. This figure presents the relative contribution of each forecast source e, given by $K_e \times contrib_e$, averaged over the 9 forecast lead times. If the information provided by each ensemble were the same, the combined relative contribution of CMC-EPS and NCEP-GEFS forecasts would have been approximately 0.44. But rather it reaches at most 0.35, which means that the members of the CMC-EPS and NCEP-GEFS ensembles are under-weighted for the benefit of the ECMWF-EPS members.

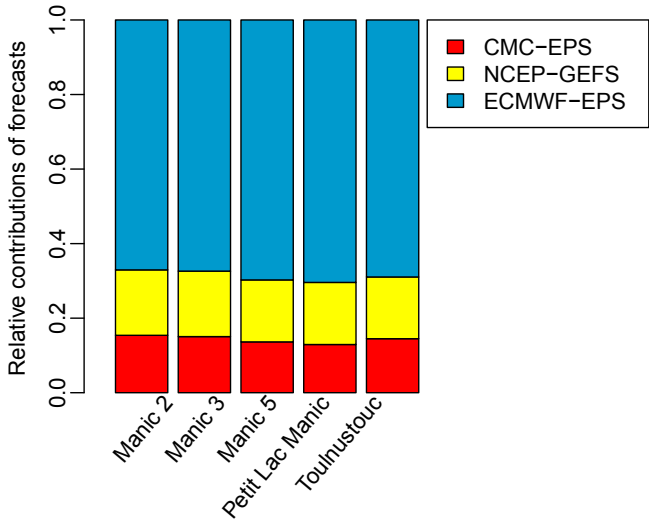


Figure 12: Relative contributions of forecast sources according to the two-variable latent model averaged over the nine forecast lead times.

6.2 Precipitation forecasts

Results on Manic 2 watershed

Returning to Manic 2 watershed as an illustrative example, we now test our post-processing method for ensemble precipitation forecasts. As for temperatures, year 2013 was employed to estimate the parameters of the post-processing Tobit model, and daily forecasts were produced for 2014.

The parameter of transformation, γ, is estimated beforehand, the value obtained is 0.43. Figure 13 shows the resulting post-processed ensemble precipitation forecasts for our illustrative example of April 30th, 2014. As for temperatures, the 90 raw scenarios have been treated independently, lead time by lead time, the post-processed outputs being re-ordered using Ensemble Copula Coupling ([Schefzik et al., 2013]). The color code used in this figure is the same as the one in Fig 9.

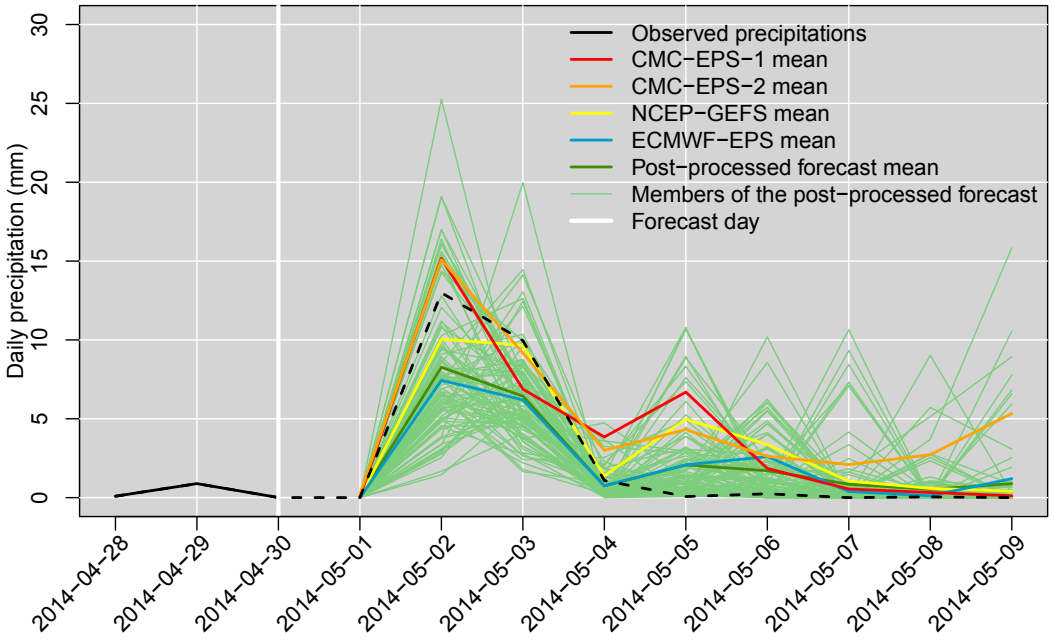


Figure 13: Example of daily precipitation forecasts for the Manic 2 watershed using the post-treatment method for NCEP-GEFS, CMC-EPS (odd : 1 and even : 2) and ECMWF-EPS ensembles. Precipitation to be forecast is in dashed black.

Based upon visual examination of rank histograms, the post-processing method seems to lead to a better reliability compared to the original overall forecasts for the Manic 2 watershed. This gain is particularly marked for the shortest forecast deadlines for which under-dispersion and bias are more pronounced. As an illustration, the rank histograms for 3-days lead time ensemble forecasts have been presented in Fig 1. We see that raw ensembles (Fig 1–(1) to Fig 1–(4)) are clearly biased and underdispersive whereas post-processed ensembles show graphically good calibration (Fig 1–(5) and Fig 1–(6)). The rank histograms corresponding to the post-processed forecasts are indeed flatter than those associated with the raw ensemble forecasts.

This gain in reliability is not reflected in the analysis of CRPS values. Figure 14 shows the performance of the forecasts as measured by this scoring rule for each lead times. It is seen, for Manic 2 watershed, that post-processing does not improve significantly the raw ensemble precipitation forecasts. However, adding more ensembles is improving forecasts. In fact, CRPS values for the grand ensemble are systematically smaller than that of the CMC-EPS members.

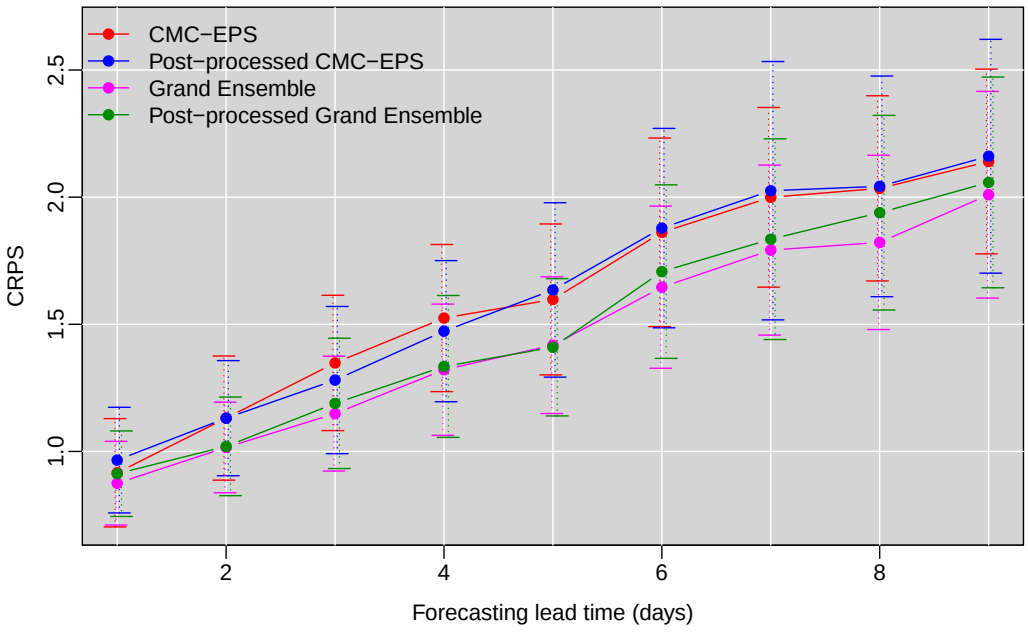


Figure 14: CRPS (in mm) for 2014 post-processed ensemble precipitation forecasts for the Manic 2 watershed as a function of the lead times and with associated bootstrap intervals. Year 2013 is used to estimate the parameters.

Results on other watersheds

For the other watersheds, the results obtained are much better: forecasts are generally slightly improved by the application of our post-treatment, particularly for small lead times, 1 up to 4 days (not shown). This is even more apparent when we consider cumulative precipitations forecasts, such as illustrated in Fig 15. In a hydrological forecasting perspective, cumulative precipitations are of particular importance since crucial decisions are taken based upon incoming inflow volumes. CMC-EPS 9-days precipitation accumulation forecasts are improved by post-processing regardless of the watershed. This is an argument in favour of a good reconstruction of the temporal dependency by the ECC-Q method, which has been used to get cumulative forecasts from the daily forecasts originally issued by our post-processing method.

The probabilistic forecasts produced using the raw large ensemble are always better than those of the CMC-EPS. On the other hand, the forecasts for the grand ensemble totals are only improved for 3 out of 5 watersheds. This can be explained by the fact that the forecasts produced using the raw large ensemble already shows a good performance on those watersheds.

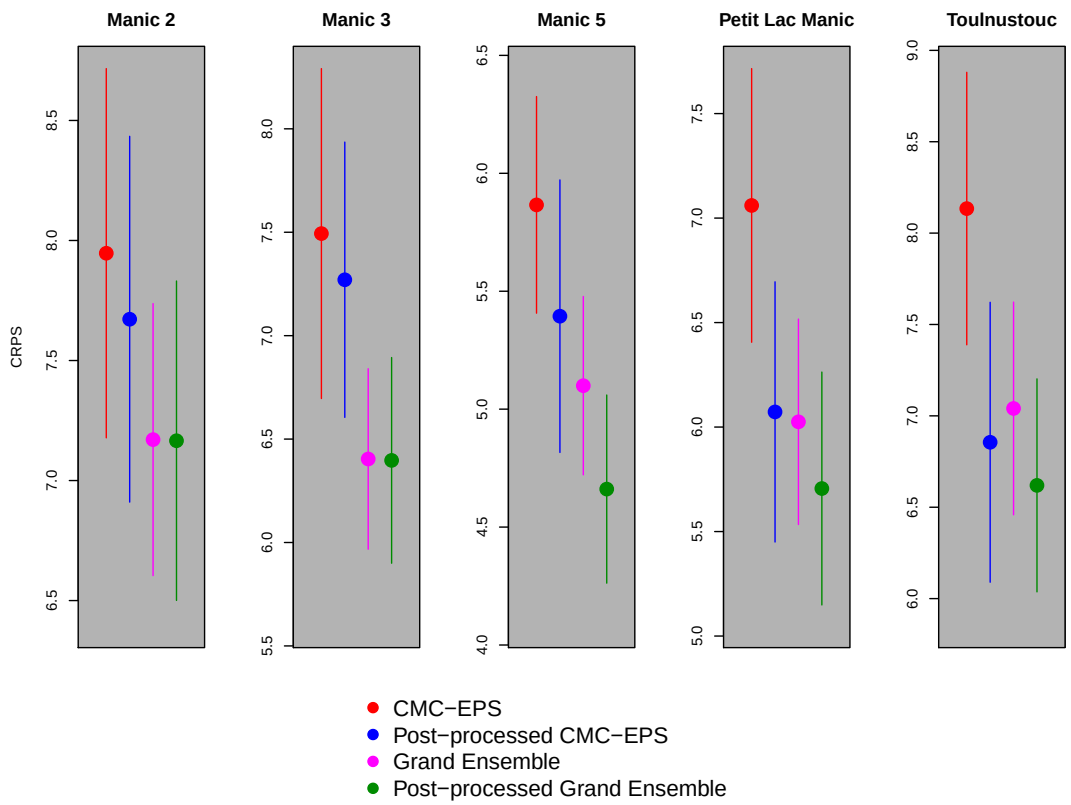


Figure 15: CRPS (in mm) for 2014 post-processed 9-days ensemble cumulative precipitation forecasts and associated bootstrap intervals. Year 2013 is used to estimate the parameters.

7 Conclusion and discussion

Based on the concept of exchangeability, a desirable property of relabeling invariance for a meteorological ensemble, we have developed in Sections 3 and 4 a constructive framework for post-processing members of multiple ensembles. This theoretically-justified mixed-effect structure gives an attractive physical interpretation to the latent variables underpinning the statistical model: they are the simple essential traits that the sophisticated numerical code mimicking the earth system model retains from the many simulations of the future weather that have been launched to give birth to the ensemble members. The model we propose also allows for the parsimonious integration of several sources of information: it can combine several ensembles produced by different meteorological centers and, eventually, deterministic weather forecasts resulting from meteorologists expertise.

We put this theoretical framework into operation for post-processing temperature forecasts, that are known for their Gaussian-like behavior as well as for precipitation forecasts, a rather more delicate statistical challenge. Due to the zero inflation of the latter distribution, we have made recourse in this case to an extension by truncation and power transform of the normal distribution. Inference methods rely on the EM and SEM algorithms to estimate the model parameters that maximize likelihood. We checked that the inference algorithms perform well on artificial data before applying the resulting post-processing methods to temperatures and precipitations data provided by Hydro-Québec.
Not surprisingly, the CRPS is almost systematically improved for all case studies after performing the post-treatments. Our statistical treatment affects both components of this scoring rule, calibration and sharpness. Calibration is improved because our fitted model has straightened the statistical features of the ensemble. Sharpness should be improved because multiple sources of information have been combined.

When adapted to precipitation ensembles by including an additional layer in the multilevel model, the method is not yet entirely satisfactory: by means of a normalized transformation of precipitation, we succeed in providing well calibrated forecasts, but their CRPS performances are a little disappointing when considering the multi-ensemble post-processing on 2 of the 5 considered watersheds, because the

forecasts obtained by considering the ensemble forecasts all together as a grand ensemble already shows a good performance on those watersheds. However, in an operational context, obtaining all the ensemble weather forecasts forming the grand ensemble in real time can be complicated. Hence the interest of a method providing satisfying performances from a subset of the ensembles involved, as the proposed method can do. Moreover, the use of a moving (and/or seasonal) training period, as did Taillardat et al. [2016] for instance, could be a way to get enhanced performances by better taking into account non-stationarities.

Inference could have been performed under the Bayesian paradigm: the Bayesian interpretation of predictive conditional distributions as a personal probabilistic judgment [Lindley, 2013] is most appropriate for operational forecasters. We nevertheless chose the EM approach for inference of the models with latent layers since Bayesian solutions often require more computational burden. In addition, the Bayesian approach really brings much improvement when the experts' beliefs can be encoding, via a proper elicitation phase, to take into account supplementary information not conveyed by the data, but this is a lengthy and difficult modeling task [O'Hagan, 2005] not developed in this paper.

Although not Bayesian, the proposed method is able to integrate expert's forecasts as a source of additional information (it suffices to consider it as a peculiar ensemble with a single member), provided some stationarity of the process of deterministic forecast expertise (the same expert or a cohesive team of forecasters) during the learning period. This happens to be the case, for example, with Hydro-Québec. This would give forecasters, wishing to keep on working with their best single estimate of the future weather, the opportunity to share their vision of the meteorologic phenomenon within a multi-ensemble aggregation process.

Post-processing the meteorological ensembles to get reliable probabilistic weather forecasts is only the beginning of the story: water in the river, not in the air, is the product of interest for hydropower companies such as HydroQuebec. As a consequence, one should also account for the uncertainty stemming from the inexact code representation of the rainfall-runoff transformation. Otherwise, some underdispersion of the probabilistic waterflow forecasts may hamper their values with regards to dam operation management. As a practical consequence, an additional statistical post-processing method is be used on the outputs of the rainfall-runoff model [Courbariaux et al., 2017].

Moreover, we have post-processed the meteorological ensembles for a single variable of interest, a chosen location and a given lead time. Here, as usual in the ensemble community, we simply rely on empirical copulas to restore coherence in space, time and between the multiple variables. Making recourse to empirical copula to retrieve a joint multivariate structure does not go without an unfortunate drawback: the number of predictive simulations cannot be greater than the size of the original ensemble. Moreover, had we post-processed an ensemble for maximum and minimum daily temperatures, the method could not guarantee that the predicted maximum daily temperatures will always remain higher than the corresponding minimum daily temperatures since these two quantities of interest would have been processed independently. A natural remedy might be to keep on taking advantage of the Gaussian properties and try to model the multivariate ensemble with many lead times and several locations as a huge Gaussian process, with an autoregressive effect in the latent variable related to the mean of the ensemble. This is of course the unattainable Grail (the inference would be much tougher), but looking forward to it can help patching additional salient traits of ensembles to fruitful variations of the simple models proposed in this paper.

8 Acknowledgements

This work was supported by Électricité de France and by Hydro-Québec [research grant number 694R] through the PhD. thesis of M. Courbariaux. We would like to thank Jacques Bernier, Joël Gailhard, Anne-Catherine Favre and Vincent Fortin for their unfailing help and constructive comments regarding this work. The forecasting and development teams at EDF-DTG and Hydro-Québec have provided the necessary material and case studies as well as many valuable advises: we thank in particular Fabian Tito Arandia Martinez and Éric Crobeddu from Hydro-Québec, Fabien Rinaldi and Rémy Garçon from EDF-DTG.

References

D. Allard. Modeling spatial and spatio-temporal non Gaussian processes. In *Advances and Challenges in Space-time Modelling of Natural Events*, pages 141–164. Springer, 2012.

Z. Ben Bouallègue. Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Weather and Forecasting*, 28(2):515–524, 2013.

P. Bougeault, Z. Toth, C. Bishop, B. Brown, D. Burridge, D. H. Chen, B. Ebert, M. Fuentes, T. M. Hamill, K. Mylne, et al. The THORPEX interactive grand global ensemble. *Bulletin of the American Meteorological Society*, 91(8):1059–1072, 2010.

G. E. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.

M. Broniatowski, G. Celeux, and J. Diebolt. Reconnaissance de mélanges de densités par un algorithme d'apprentissage probabiliste. *Data Analysis and Informatics*, 3:359–373, 1983.

R. Buizza, M. Leutbecher, and L. Isaksen. Potential use of an ensemble of analyses in the ECMWF Ensemble Prediction System. *Quarterly Journal of the Royal Meteorological Society*, 134(637):2051–2066, 2008.

G. Celeux and J. Diebolt. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2(1):73–82, 1985.

M. Courbariaux, P. Barbillon, and É. Parent. Water flow probabilistic predictions based on a rainfall–runoff simulator: a two-regime model with variable selection. *Journal of Agricultural, Biological and Environmental Statistics*, 22(2):194–219, 2017.

B. de Finetti. Funzione caratteristica di un fenomeno aleatorio. 1931.

B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68, 1937.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.

C. Fraley, A. E. Raftery, and T. Gneiting. Calibrating multimodel forecast ensembles with exchangeable and missing members using bayesian model averaging. *Monthly Weather Review*, 138(1):190–202, 2010.

R. Garçon. Prévision opérationnelle des apports de la Durance à Serre-Ponçon à l'aide du modèle MORDOR. Bilan de l'année 1994-1995. *La Houille Blanche*, (5):71–76, 1996.

A. E. Gelfand and A. F. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409, 1990.

T. Gneiting, A. E. Raftery, A. H. Westveld, and T. Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5): 1098–1118, 2005.

T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

C. Guay, M. Minville, and I. Chartier. Hsami+ : Guide théorique. Technical report, Institut de recherche d'Hydro-Québec, Varennes, QC, Canada, 2018.

T. M. Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3):550–560, 2001.

T. M. Hamill and S. J. Colucci. Verification of eta-rsm short-range ensemble forecasts. *Monthly Weather Review*, 125(6):1312–1327, 1997.

H. Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000.

E. Hewitt and L. J. Savage. Symmetric measures on Cartesian products. *Transactions of the American Mathematical Society*, 80(2):470–501, 1955.

S. Khajehei, A. Ahmadalipour, and H. Moradkhani. An effective post-processing of the north american multi-model ensemble (nmme) precipitation forecasts over the continental us. *Climate Dynamics*, 51 (1-2):457–472, 2018.

R. Krzysztofowicz and C. J. Maranzano. Bayesian processor of output for probabilistic quantitative precipitation forecasts. *Manuscript in review*, 2006.

W. Li, Q. Duan, C. Miao, A. Ye, W. Gong, and Z. Di. A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *Wiley Interdisciplinary Reviews: Water*, 4(6):e1246, 2017.

D. V. Lindley. *Understanding uncertainty*. John Wiley & Sons, 2013.

N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

J. W. Messner, G. J. Mayr, A. Zeileis, and D. S. Wilks. Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Monthly Weather Review*, 142(1):448–456, 2014.

A. O'Hagan. *Research in elicitation*. University of Sheffield, Department of Probability and Statistics, School of Mathematics, 2005.

Y.-Y. Park, R. Buizza, and M. Leutbecher. Tigge: Preliminary results on comparing and combining ensembles. *Quarterly Journal of the Royal Meteorological Society*, 134(637):2029–2050, 2008. ISSN 1477-870X. doi: 10.1002/qj.334. URL `http://dx.doi.org/10.1002/qj.334`.

L. Perreault. Post-traitement statistique des prévisions météorologiques d'ensemble pour le complexe manicouagan : les températures. Rapport scientifique IREQ-2017-0057, Institut de recherche d'Hydro-Québec, 2017.

A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5), 2005.

R. Schefzik, T. L. Thorarinsdottir, T. Gneiting, et al. Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28(4):616–640, 2013.

M. Scheuerer. Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140(680):1086–1096, 2014.

M. Scheuerer and T. M. Hamill. Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, 143(11):4578–4596, 2015.

P. Schultz, H. Yuan, M. Charles, R. Krzysztofowicz, and Z. Toth. Pseudo-precipitation: a continuous variable for statistical post-processing. In *20th Conference on Probability and Statistics in the Atmospheric Sciences*, 2010.

J. M. L. Sloughter, A. E. Raftery, T. Gneiting, and C. Fraley. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review*, 135(9):3209–3220, 2007.

M. Taillardat, O. Mestre, M. Zamo, and P. Naveau. Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144(6):2375–2393, 2016.

C. Tebaldi and R. Knutti. The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 365(1857):2053–2075, 2007.

T. L. Thorarinsdottir and T. Gneiting. Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2):371–388, 2010.

D. S. Wilks. Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications*, 16(3):361–368, 2009.

A Details on inference in the Gaussian case

E-step We need to compute the conditional distributions $\left[Z_t|\omega_t^{-2}, \mathbf{X}_t, Y_t\right]$ and $\left[\omega_t^{-2}|\mathbf{X}_t, Y_t\right]$ for each time t of the training set. We interpret the joint distribution at time t

$$\begin{aligned}\left[Z_t, \omega_t^2, \mathbf{X}_t, Y_t\right] &= [\mathbf{X}_t, Y_t|Z_t, \omega_t^{-2}][\omega_t^{-2}, Z_t] \\ &= [\mathbf{X}_t|Z_t, \omega_t^{-2}][Y_t|Z_t, \omega_t^{-2}][\omega_t^{-2}, Z_t]\end{aligned}$$

as a function of $\left(\omega_t^{-2}, Z_t\right)$, and try to recognize the probability distribution function (pdf) of $\left(Z_t, \omega_t^{-2}|\mathbf{X}_t, Y_t\right)$ up to a multiplicative constant, since $\left[Z_t, \omega_t^{-2}|\mathbf{X}_t, Y_t\right] = [\omega_t^{-2}, Z_t, \mathbf{X}_t, Y_t] \times \left(\frac{1}{[\mathbf{X}_t, Y_t]}\right)$).

The complete deviance (minus twice the complete loglikelihood) at time t can be written as a quadratic form in $\mathbf{X}_t$ and Y_t, up to known normalizing constants:

$$\begin{aligned}\sum_{e=0}^{E}\left\{\sum_{k=1}^{K_e}(X_{e,k,t} - b_e Z_t - a_e)^2\omega_t^{-2}c_e^{-2} - K_e\log\left(\omega_t^{-2}\right) - K_e\log\left(c_e^{-2}\right)\right\} \\ + Z_t^2\lambda^{-1}\omega_t^{-2} - \log(\lambda^{-1}) - \log\left(\omega_t^{-2}\right) + 2\beta\omega_t^{-2} - 2(\alpha-1)\log\left(\omega_t^{-2}\right) - 2\log\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right). \quad (8)\end{aligned}$$

The pdf we are looking for can be further decomposed as:

$$\left[Z_t, \omega_t^{-2}|\mathbf{X}_t, Y_t\right] = [Z_t|\omega_t^{-2}, \mathbf{X}_t, Y_t][\omega_t^{-2}|\mathbf{X}_t, Y_t]\,.$$

We now check if we can still benefit from a conjugate situation, i.e. given $(\mathbf{X}_t, Y_t)$ we would still get a normal pdf for $\left(Z_t|\omega_t^{-2}, \mathbf{X}_t, Y_t\right)$ under the form $\mathcal{N}(m', \lambda'\omega_t^2)$ and a gamma pdf for $\left(\omega_t^{-2}|\mathbf{X}_t, Y_t\right)$, $\Gamma(\alpha', \beta')$. Expressed as a function of $[Z_t, \omega_t^{-2}|\mathbf{X}_t, Y_t]$, the deviance exhibits the following shape:

$$(Z_t - m_t')^2\lambda'^{-1}\omega_t^{-2} - \log(\omega_t^{-2}) - \log(\lambda'^{-1}) - 2(\alpha'-1)\log(\omega_t^{-2}) + 2\beta_t'(\omega_t^{-2})\,.$$

We proceed by trying to identify parameters $\alpha', \beta_t', \lambda', m_t'$ in the above equation to match the deviance for their joint distribution given by Eq (8).

By matching both expressions, we obtain:

$$\begin{aligned}\lambda'^{-1} &= \textstyle\sum_{e=0}^{E} K_e b_e^2 c_e^{-2} + \lambda^{-1}, \\ m_t' &= \lambda' \cdot \textstyle\sum_{e=0}^{E} c_e^{-2} b_e K_e\left(\bar{X}_{e,t} - a_e\right) \\ \alpha' &= \alpha + \frac{\sum_{e=0}^{E} K_e}{2}, \\ \beta_t' &= \beta + \tfrac{1}{2}\left\{\textstyle\sum_{e=0}^{E}\sum_{k=1}^{K_e} c_e^{-2}(X_{e,k,t} - a_e)^2 - m_t'^2\lambda'^{-1}\right\},\end{aligned}$$

where $\bar{X}_{e,t} = \frac{1}{K_e}\sum_{k=1}^{K_e} X_{e,k,t}$. Therefore, we are in a conjugate situation since the conditional pdf $\left[Z_t, \omega_t^{-2}|\mathbf{X}_t, Y_t\right]$ is in the normal-gamma model as is the marginal pdf $\left[Z_t, \omega_t^{-2}\right]$.

Denoting $\phi(\cdot)$ the first derivative of function $\log\{\Gamma(\cdot)\}$, the moments necessary for performing the E-step are:

$$\begin{aligned}\mathbb{E}\left(\log(\omega_t^{-2})|\mathbf{X}_t, Y_t\right) &= -\log(\beta_t') + \phi(\alpha')\,, \\ \mathbb{E}(\omega_t^{-2}|\mathbf{X}_t, Y_t) &= \tfrac{\alpha'}{\beta_t'}\,, \\ \mathbb{E}(Z_t^2\omega_t^{-2}|\mathbf{X}_t, Y_t) &= \lambda' + m_t'^2\tfrac{\alpha'}{\beta_t'}\,, \\ \mathbb{E}(Z_t\omega_t^{-2}|\mathbf{X}_t, Y_t) &= m_t'\tfrac{\alpha'}{\beta_t'}\,.\end{aligned}$$

M-step We write the complete deviance $D(\boldsymbol{\theta}) = D(\alpha, \beta, \lambda, \mathbf{a}, \mathbf{b}, \mathbf{c})$, denoting n the number of records in the data set (each of them indexed by t):

$$\begin{aligned}D(\boldsymbol{\theta}) = \sum_{t=1}^{n}\Bigg\{ & Z_t^2\lambda^{-1}\omega_t^{-2} - \log(\lambda^{-1}) - 2\alpha\log(\omega_t^{-2}) + 2\beta\omega_t^{-2} - 2\alpha\log(\beta) \\ & + 2\log\{\Gamma(\alpha)\} \\ & + \sum_{e=0}^{E}\left\{\sum_{k=1}^{K_e}\left(X_{e,k,t} - a_e - b_e Z_t\right)^2 c_e^{-2}\omega_t^{-2} - \log\left(c_e^{-2}\right)\right\}\Bigg\} + \text{Cst}\,,\end{aligned}$$

where Cst is a constant term with respect to the parameters to estimate.

First, the expectation of $D(\boldsymbol{\theta})$ is computed by using the moments computed in the E-step. Then, this expectation is differentiated with respect to the parameters to be updated. This leads to the following explicit update formulas, the subscript *new* indicates the new value of the parameter:

$$
\begin{array}{lcll}
b_{e,new} & = & \frac{\frac{D_e}{B} - \frac{C_e}{G}}{\frac{G}{B} - \frac{H}{G} - \frac{n\lambda'}{G\alpha'}} & \text{for } e \in \{1, \ldots, E\}, \\
a_{e,new} & = & \frac{D_e}{B} - b_{e,new}\frac{G}{B} & \text{for } e \in \{0, \ldots, E\}, \\
c^2_{e,new} & = & K_e b^2_{e,new}\lambda' + \frac{1}{n}\sum_{t=1}^{n}\frac{\alpha'}{\beta'_t}\sum_{k=1}^{K_e}\left(X_{e,k,t} - a_{e,new} - b_{e,new}m'_t\right)^2 & \text{for } e \in \{1, \ldots, E\}, \\
\lambda_{new} & = & \lambda' + \frac{\alpha'}{n}\sum_{t=1}^{n}\frac{m'^2_t}{\beta'_t}, & \\
\beta_{new} & = & \frac{n\alpha_{new}}{\alpha'\sum_{t=1}^{n}\frac{1}{\beta'_t}}, &
\end{array}
$$

where $G = \sum_{t=1}^{n}\frac{m'_t}{\beta'_t}$, $B = \sum_{t=1}^{n}\frac{1}{\beta'_t}$, $C_e = \sum_{t=1}^{n}\frac{m'_t\bar{X}_{e,t}}{\beta'_t}$, $D_e = \sum_{t=1}^{n}\frac{\bar{X}_{e,t}}{\beta'_t}$ et $H = \sum_{t=1}^{n}\frac{m'^2_t}{\beta'_t}$. For updating α, we use a numeric solver of the following equation:

$$
\log\left(\frac{n\alpha_{new}}{\alpha'\sum_{t=1}^{n}\frac{1}{\beta'_t}}\right) - \phi(\alpha_{new}) = \frac{1}{n}\sum_{t=1}^{n}\log(\beta'_t) - \phi(\alpha').
$$