RMetS
Royal Meteorological Society

On the calibration of multilevel Monte Carlo ensemble forecasts

A. Gregory* and C. J. Cotter
Department of Mathematics, Imperial College London, UK

*Correspondence to: A. Gregory, Department of Mathematics, Imperial College London, Kensington, London SW7 2AZ, UK.
E-mail: a.gregory14@imperial.ac.uk

The multilevel Monte Carlo method can efficiently compute statistical estimates of discretized random variables for a given error tolerance. Traditionally, only a certain statistic is computed from a particular implementation of multilevel Monte Carlo. This article considers the multilevel case in which one wants to verify and evaluate a single ensemble that forms an empirical approximation to many different statistics, namely an ensemble forecast. We propose a simple algorithm that, in the univariate case, allows one to derive a statistically consistent single ensemble forecast from the hierarchy of ensembles that are formed during an implementation of multilevel Monte Carlo. This ensemble forecast then allows the entire multilevel hierarchy of ensembles to be evaluated using standard ensemble forecast verification techniques. We demonstrate the case of evaluating the calibration of the forecast.

Key Words: multilevel Monte Carlo; ensemble forecasts; rank histograms

Received 7 November 2016; Revised 28 March 2017; Accepted 29 March 2017; Published online in Wiley Online Library 23 May 2017

1. Introduction

Multilevel Monte Carlo (MLMC: Giles, 2008) is a technique that has gained significant popularity over the past decade. It is designed to produce statistical estimators for discretized random variables at significantly lower computational costs than their Monte Carlo counterparts for a fixed error. This is done by using a hierarchy of larger ensembles using lower accuracy models and smaller ensembles using higher accuracy models. For probabilistic forecasting, one can use this multilevel technique to estimate statistics from a forecast probability distribution, given some distribution of the initial conditions and/or random forcing.

In the multilevel Monte Carlo framework, one usually considers a particular statistic, such as evaluations of the cumulative distribution function (CDF: Elfverson *et al.*, 2014; Giles *et al.*, 2015; Wilson and Baker, 2016), probability density function (PDF: Bierig and Chernov, 2016) or expected values (Giles, 2008; Cliffe *et al.*, 2011), selecting the ensemble sizes/finest level of resolution so that the overall multilevel estimator produces an efficient and accurate approximation.

In the case of ensemble forecasting, one usually wishes to compute many statistics from the same ensemble. These approximations can be assessed using suitable verification techniques. Verification tools used within ensemble forecasting usually work alongside observations of the process that one is interested in forecasting and can help verify properties from calibration to the sharpness of a forecast (Gneiting *et al.*, 2007).

Given the multilevel hierarchy of ensembles from different resolutions that form MLMC estimates of statistics, we would also like to evaluate/verify these ensembles in the same way; this is the subject of this article. We propose a methodology to take observables of a univariate random variable, or scalar observables of a multidimensional random variable (such as a random field evaluated at a point in space), from a multilevel hierarchy of ensembles with varying resolutions and generate an accompanying single ensemble forecast. Most of the standard techniques in the field of ensemble forecasting are limited to the univariate case; in the context of large dimensional models in weather and climate, these are usually applied to scalar observables such as point values or integral quantities.

This single forecast is statistically consistent with the multilevel estimate. It can then be used to verify the forecast from the original ensemble hierarchy using standard methods such as calibration tests.

An alternative approach to this could be approximating each verification or scoring measure, such as the calibration or sharpness, individually and directly from the multilevel hierarchy of ensembles. For example, one could use a MLMC approximation for the CDF (Elfverson *et al.*, 2014; Giles *et al.*, 2015) to help compute a rank histogram to evaluate the forecast calibration. Each different MLMC approximation typically comes with a framework to implement it, such as a smoothing scheme in the former of those two studies.

However, by using the proposed methodology in this article, one does not need a different multilevel approximation and framework for each individual scoring measure; instead, any standard verification technique, such as the calibration or continuous ranked probability score (CRPS: Gneiting *et al.*, 2007), can be employed on this standard single ensemble forecast.

To generate this ensemble forecast, inverse-transform sampling is used. The new single ensemble forecast preserves the unbiased approximation to the mean of the forecast distribution from the original multilevel estimator and forms a consistent approximations to other statistics, such as higher moments.

This study proceeds as follows; an introduction to MLMC will be given in section 2, then a simple method to find a consistent ensemble forecast from a MLMC approximation will be given in section 3, alongside a corresponding verification technique for these ensemble forecasts. Finally, a conclusion follows.

2. Multilevel Monte Carlo

Multilevel Monte Carlo (Giles, 2008) is used primarily as a computationally cheap alternative to an equivalent accuracy single-level Monte Carlo estimator of statistics with respect to a probability distribution. Suppose one wishes to compute estimates to statistics of $f(X_{L,t})$, such as $\mathbb{E}[f(X_{L,t})]$, where $X_{L,t}$ is a numerical approximation of our 'forecast' random variable X (with discretization parameter $h_L \propto M^{-L}$, $M > 1$) at time $t \geq 0$ and f is some scalar observable function. Let $X^i_{L,t}$, $i = 1, \ldots, N$, be $N \geq 1$ independent and identically distributed (i.i.d.) samples of the random variable $X_{L,t}$. Then an empirical approximation to the density of $X_{L,t}$ is

$$\pi^{MC}_{L,t}(x) = \frac{1}{N}\sum_{i=1}^{N}\delta(x - X^i_{L,t}), \tag{1}$$

where δ is the Dirac delta function. One can then estimate statistics for this empirical distribution via the Monte Carlo method. For example, the standard estimator for $\mathbb{E}[f(X_{L,t})]$ is given by

$$\bar{f}^{MC}_{L,t} = \frac{1}{N}\sum_{i=1}^{N} f(X^i_{L,t}). \tag{2}$$

Now consider the multilevel framework, using $L + 1$ ensembles $\{X^i_{l-1,t}, X^i_{l,t}\}^{l=L,i=N_l}_{l=0,i=1}$ (with $X^i_{-1} = 0$) of sizes N_l, to derive the equivalent MLMC approximation to $\mathbb{E}[f(X_{L,t})]$,

$$\bar{f}_{L,t} = \frac{1}{N_0}\sum_{i=1}^{N_0} f(X^i_{0,t}) + \sum_{l=1}^{L}\left(\frac{1}{N_l}\sum_{i=1}^{N_l}[f(X^i_{l,t}) - f(X^i_{l-1,t})]\right). \tag{3}$$

Taking the telescoping sum of expectations,

$$\mathbb{E}[f(X_{L,t})] = \mathbb{E}[f(X_{0,t})] + \sum_{l=1}^{L}\mathbb{E}[f(X_{l,t})] - \mathbb{E}[f(X_{l-1,t})] \tag{4}$$

and considering

$$\mathbb{E}[\hat{f}_{l,t}] = \begin{cases} \mathbb{E}[f(X_{0,t})], & l = 0, \\ \mathbb{E}[f(X_{l,t})] - \mathbb{E}[f(X_{l-1,t})], & l > 0, \end{cases} \tag{5}$$

where

$$\hat{f}_{l,t} = \begin{cases} \sum_{i=1}^{N_0} \frac{1}{N_0} f(X^i_{0,t}), & l = 0, \\ \sum_{i=1}^{N_l} \frac{1}{N_l}\big(f(X^i_{l,t}) - f(X^i_{l-1,t})\big), & l > 0, \end{cases} \tag{6}$$

one recovers $\bar{f}_{L,t}$ as an unbiased approximation of $\mathbb{E}[f(X_{L,t})]$. The important thing to note here is that the fine (level l) and coarse (level $l - 1$) samples in the difference estimators, $\hat{f}_{l,t}$, must be positively correlated for each i. This can be achieved by using the same random system input (e.g. initial conditions/stochastic forcing) for each i on both levels. On the other hand, the samples in different ensembles must be uncorrelated. The uses of the above framework are incredibly varied. One can even condition these multilevel estimators on observations using processes such as filtering (Jasra *et al.*, 2015; Gregory *et al.*, 2016; Gregory and Cotter, 2016). In addition to this, there have been many other applications of MLMC, some of which are highlighted in the review of Giles (2015).

Given an optimal choice of L and N_l, one can compute these estimators with the same accuracy as their standard Monte Carlo counterparts for significantly less computational expense. This works by noting that, due to the correlation between the pairs of samples in each difference estimator, the sample variance of $f(X_{l,t}) - f(X_{l-1,t})$, denoted V_l, should decrease asymptotically with $l \to \infty$. If one desires the accuracy of $\bar{f}_{L,t}$ to be

$$\mathbb{E}\left(\left(\bar{f}_{L,t} - \mathbb{E}[f_{L,t}]\right)^2\right) < \epsilon^2, \tag{7}$$

then one can follow the algorithm in Giles (2008) to compute $\bar{f}_{L,t}$ by updating, online (as one adds additional samples), the optimal sample sizes

$$N_l = \left\lceil 2\epsilon^{-2}\,(V_l h_l)\left(\sum_{n=0}^{L}\sqrt{V_n/h_n}\right)\right\rceil, \tag{8}$$

whilst increasing L until

$$|\hat{f}_{L,t}| < \frac{1}{\sqrt{2}}(M-1)\epsilon.$$

An estimated V_l can be used in the optimal sample size formula. Computational cost reductions occur because, if V_l decreases asymptotically with $l \to \infty$, then N_l does also, leading to a trade-off between estimator variance and bias in each difference estimator. To conclude, we should have large ensembles for the lower levels and smaller ensembles on the higher levels, given by asymptotically decreasing values of V_l.

For the full algorithm and corresponding theory, see Giles (2008, 2015).

3. Ensemble forecasting

This article now proposes a method to generate a single ensemble forecast from the hierarchy of ensembles created from the MLMC method. Put simply, one can generate a large ensemble (much larger than the finest level ensemble) that represents the entire MLMC approximation to the forecast distribution. This is more useful for verification of the hierarchy of ensembles, rather than simply using standard verification techniques on the finest ensemble in this hierarchy. As mentioned in the previous section, the sample sizes, N_l, for pairs of ensembles on all levels decrease asymptotically and thus the finest ensemble is the smallest ensemble in the hierarchy. Using the finest ensemble for verification of the entire MLMC approximation of the forecast distribution would neglect the majority of samples, on lower levels, from which the approximation was composed.

In addition to this verification, given the statistical consistency of this ensemble forecast with the multilevel ensemble hierarchy, many statistics can be easily estimated via this ensemble.

3.1. *Multilevel Monte Carlo ensemble forecasts*

Now, assume $X_{l,t} \in \mathbb{R}$ and so, if $X_{l,t}$ was multivariate in the section before, $X_{l,t}$ now represents $f(X_{l,t})$, the scalar observable, e,g, $X_{l,t}$ evaluated at a point in space. Here we describe how to generate the single ensemble forecast of a scalar observable $X^i_{F,t} \in \mathbb{R}$, $i = 1, \ldots, N$ from the MLMC hierarchy of ensembles through inverse-transform sampling. It is important to note that this ensemble does not contain i.i.d. samples from the forecast distribution; instead they will simply be approximations to these samples. However, this single ensemble has the properties to form a consistent empirical estimate to the forecast distribution and associated distribution functions.

From here onwards, we will assume that values of N_l and L have been either set or found and that the hierarchy of ensembles $\{X^i_{l-1,t}, X^i_{l,t}\}^{l=L,i=N_l}_{l=0,i=1}$, with $X_{-1,t} = 0$, has been generated. Predominantly, this is because the framework that this article presents is designed for evaluating any given MLMC approximation. Each approximation has a hierarchy of ensembles that use values of N_l and L that have been optimized around minimizing the cost of that particular approximation. Each approximation typically comes with its own algorithm to set up these values. Thus, by making the aforementioned assumption we can keep this framework general to all approximations. In addition to this, it is likely that in real forecasting practice one would pick the desired maximum level L and then set fixed values of N_l based on the maximum computational expense one can use on a particular level. This way of choosing N_l and L is implemented in the numerical example later in the article.

Inverse transform sampling is the process of evaluating an (approximation to the) inverse CDF, $F^{-1}(u)$, $u \in [0,1]$, also known as the *quantile function*. In the case where the CDF, F, of a random variable is strictly increasing and absolutely continuous, there exists a unique value $x \equiv F^{-1}(u)$ for which $F(x) = u$. This distribution must usually be estimated empirically. If the true CDF of the forecast distribution is known to be absolutely continuous and the samples are sorted to form order statistics, then some of these estimates have been shown to be consistent approximations to $F^{-1}(u)$ (Ma *et al.*, 2011). A very simple consistent estimate for an evaluation to the quantile function of the distribution with CDF, F, using the (ascending) sorted samples $\{X^i\}_{i=1,...,N} \sim F$, $X^1 < X^2 < .... < X^N$ is

$$\hat{F}^{-1}(u) = X^{\lceil N\times u\rceil}. \tag{9}$$

Here, the estimate is a consistent one in the sense that it converges in probability to $F^{-1}(u)$ as $N \rightarrow \infty$. One can use linear interpolation and extrapolation to smooth this consistent estimate. Other inconsistent techniques include fitting a parametric distribution to the ensemble, such as a Gaussian, and sampling from a closed-form quantile function (e.g. Φ for a Gaussian distribution) for that distribution. In all cases, when the empirical quantile function is evaluated with i.i.d. uniform samples $u \in [0,1]$, approximations to samples of X can be generated.

The use of inverse-transform sampling alongside MLMC was first suggested in Giles (2013), who proposed to use it to minimize the discrete Wasserstein distance between the two paired ensembles in each difference estimator within (3) and thus positively couple them. Instead, here we will use inverse-transform sampling in the context of a MLMC approximation to the quantile function of the forecast distribution,

$$\bar{F}^{-1}_{L,t}(u) = R(X)^{\lceil N_0\times u\rceil}_{0,t} + \sum_{l=1}^{L}\left(R(X)^{\lceil N_l\times u\rceil}_{l,t} - R(X)^{\lceil N_l\times u\rceil}_{l-1,t}\right), \tag{10}$$

where $R(X)^i_l$ is the ith order statistic of X_l, so that $R(X)^1_l < R(X)^2_l < ... < R(X)^{N_l}_l$. Note that there is not an exact cancellation in expected values of the above estimator terms, as in the telescoping sum of expectations in (4), as the individual approximations on each level are not unbiased, only consistent in the limit of $N_l \rightarrow \infty$. The algorithm in Table 1 demonstrates how to generate an ensemble $\{X^i_{F,t}\}_{i=1,...,N}$ of arbitrary size N, approximating samples of $X_{L,t}$.

Table 1. Algorithm demonstrating how to generate an ensemble $\{X^i_{F,t}\}_{i=1,...,N}$ of arbitrary size N, approximating samples of $X_{L,t}$.

```
procedure
    for $l = 0, ..., L$ do
        if $l = 0$ then
            Sort $X^j_{0,t}$, $j = 1, ..., N_0$, so that $R(X)^1_{0,t} < R(X)^2_{0,t} < ... < R(X)^{N_0}_{0,t}$
        else
            Sort $X^j_{l,t}$, $X^j_{l-1,t}$, $j = 1, ..., N_l$, so that $R(X)^1_{l,t} < R(X)^2_{l,t} < ... < R(X)^{N_l}_{l,t}$ and $R(X)^1_{l-1,t} < R(X)^2_{l-1,t} < ... < R(X)^{N_l}_{l-1,t}$
        end if
    end for
    for $i = 1, ..., N$ do
        Set $X^i_{F,t} = 0$
        Sample $u^i \sim U[0,1]$
        for $l = 0, ..., L$ do
            if $l = 0$ then
                $X^i_{F,t} += R(X)^{\lceil N_0\times u^i\rceil}_{0,t}$
            else
                $X^i_{F,t} += R(X)^{\lceil N_l\times u^i\rceil}_{l,t} - R(X)^{\lceil N_l\times u^i\rceil}_{l-1,t}$
            end if
        end for
    end for
end procedure
```

Note that these $X^i_{F,t}$ are not samples from $X_{L,t}$; they are only consistent approximations to the evaluations of $F^{-1}_{L,t}(u)$ for a particular u. More specifically, for a random uniform sample $u \sim U[0,1]$, we have

$$x = \bar{F}^{-1}_{L,t}(u) \tag{11}$$

and, as $N_l \rightarrow \infty$ for all l,

$$x \xrightarrow{p} F^{-1}_{L,t}(u), \tag{12}$$

where N_l are the number of samples used in each difference estimator in (10). Then, in this limit, x converges in probability to a sample from the forecast distribution on the finest level, i.e. $x \sim X_{L,t}$. Therefore any statistical estimate using these samples is a consistent one within this limit.

The single ensemble $\{X^i_{F,t}\}_{i=1,...,N}$ can form valid and consistent approximations to statistics of the forecast distribution. For example, the empirical, consistent, CDF of this ensemble forecast found from the MLMC approximation to the forecast distribution is

$$\hat{F}_{X_{F,t}}(x) = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}_{X^i_{F,t}\leq x}, \tag{13}$$

where $\mathbb{I}$ is the indicator function. Clearly this is non-decreasing for continuous $X_{F,t}$ and has the support of $[0,1]$.

One assumes that, in practice, the computational effort of evaluating the above function a large number of times to generate the ensemble $\{X^i_{F,t}\}_{i=1,...,N}$ is negligible in comparison with the expense of generating the original samples on all of the different levels. Thus, the method seems likely to be admissible even when N is much larger than N_0. Having said this, it makes sense here to set $N \propto N_0$, so that both aspects of the approximation (inverse CDF estimator and the ensemble forecast) converge in probability simultaneously. We take $N = \alpha N_0$ with $\alpha \in \mathbb{Z}, \alpha \geq 1$ for simplicity.

The proposed ensemble forecast also preserves the unbiasedness of the approximation to the first moment of the forecast distribution from the original MLMC approximation. To show this, let $\bar{X}_{F,t} = \frac{1}{\alpha N_0}\sum_{i=1}^{\alpha N_0} X^i_{F,t}$ be the sample mean of the ensemble forecast from the multilevel hierarchy of ensembles. Then

$$\begin{aligned}
\bar{X}_{F,t} =& \frac{1}{\alpha N_0}\sum_{i=1}^{\alpha N_0} X^{i}_{F,t}\\
=& \Big(\frac{1}{\alpha N_0}\sum_{i=1}^{\alpha N_0} \hat{F}_{0,t}^{-1}(u^i)\Big)\\
&+\sum_{l=1}^{L}\Bigg(\Big(\frac{1}{\alpha N_0}\sum_{i=1}^{\alpha N_0}\hat{F}_{l,t}^{-1}(u^i)\Big)-\Big(\frac{1}{\alpha N_0}\sum_{i=1}^{\alpha N_0}\hat{F}_{l-1,t}^{-1}(u^i)\Big)\Bigg),\\
=& \Big(\frac{1}{\alpha N_0}\sum_{i=1}^{\alpha N_0} X_{0,t}^{\lceil N_0\times u^i\rceil}\Big)\\
&+\sum_{l=1}^{L}\Bigg(\frac{1}{\alpha N_0}\sum_{i=1}^{\alpha N_0}\Big(X_{l,t}^{\lceil N_l\times u^i\rceil}-X_{l-1,t}^{\lceil N_l\times u^i\rceil}\Big)\Bigg)
\end{aligned} \tag{14}$$

and, given that u^i are i.i.d. draws of the uniform distribution $Unif[0,1]$, $i=1,\dots,\alpha N_0$, then

$$\begin{aligned}
\mathbb{E}[\bar{X}_{F,t}] =& \Big(\frac{1}{\alpha N_0}\sum_{i=1}^{\alpha N_0}\mathbb{E}[X_{0,t}]\Big)\\
&+\sum_{l=1}^{L}\Bigg(\frac{1}{\alpha N_0}\sum_{i=1}^{\alpha N_0}\Big(\mathbb{E}[X_{l,t}]-\mathbb{E}[X_{l-1,t}]\Big)\Bigg) \qquad (15)\\
=& \mathbb{E}[X_{0,t}]+\sum_{l=1}^{L}\mathbb{E}[X_{l,t}-X_{l-1,t}]=\mathbb{E}[X_{L,t}].
\end{aligned}$$

3.2. *Assessing the calibration of multilevel Monte Carlo ensemble forecasts*

Evaluating the ensembles used in ensemble forecasts is very important in checking the predictive value of the forecast. The remainder of this article concentrates on a method of evaluating the calibration of forecasts from the MLMC approximations to the forecast distribution directly, via the single ensemble forecast found in the previous section: the Probability Integral Transform Histogram. This technique uses observations from the target distribution to evaluate ensemble forecasts. Calibration is the measure of whether the observations are indistinguishable from the samples of the ensemble forecast distribution (Carney and Cunningham, 2006). This is a quality of the empirical forecast distribution that is possibly disregarded if one were simply to study errors of point statistical estimators.

Consider the target distribution, Y_{obs,t_k}, behind the observed process, where partial observations y_{obs,t_k}, are taken from a single realization of this process at times t_k, $k\in[0,N_y]$, $t_0=0$, $t_{N_y}=T$. Clearly, our aim would be to use a forecast distribution associated with the random variable $X_{t_k}=Y_{\mathrm{obs},t_k}$, however in many real-world scenarios Y_{obs,t_k} is unknown. Therefore verification techniques are used to rank forecasts on their similarity to the observed process, with the aim of finding the best forecast/model that derived them. The case of $X_{t_k}=Y_{\mathrm{obs},t_k}$ is known as the random variable with associated forecast distribution from the perfect model.

3.2.1. *Probability integral transform histogram*

The Probability Integral Transform (PIT) histogram is used to determine the uniformity of the observations with respect to the (empirical) CDF of the ensemble and thus the calibration of the forecast distribution with respect to the target distribution. One can define a random variable $R\sim F_{L,t_k}(Y_{\mathrm{obs},t_k})$, the PIT. Then samples of R are given by

$$r_{t_k}=F_{L,t_k}(y_{\mathrm{obs},t_k}). \tag{16}$$

The forecast distribution is said to be calibrated with respect to the target distribution if $R\sim Unif[0,1]$ and so a histogram of r_{t_k} would be relatively flat. Using the MLMC approximation to the forecast distribution, define the associated multilevel empirical PIT samples as

$$\hat{r}_{t_k}=\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}_{X^{i}_{F,t_k}\le y_{\mathrm{obs},t_k}}, \tag{17}$$

where X^i_{F,t_k} are an arbitrary N members of the ensemble forecast from the multilevel hierarchy of ensembles at time t_k using the aforementioned inverse-transform sampling method. This is simply the empirical cumulative distribution function of the N ensemble forecast members X^i_{F,t_k}. Here, given that we set $N\propto N_0$, then, in the limit of $N_l\to\infty$, for all $l=0,\dots,L$,

$$X_{F,t_k}\sim F^{-1}_{L,t_k} \tag{18}$$

and thus $F_{L,t_k}(X_{F,t_k})\sim U[0,1]$. By considering this, we have a consistent estimate of the PIT sample r_{t_k}, when concentrating on the limit of $N\propto N_0\to\infty$. One can find the frequency (H_i, $i=1,\dots,B$) of B evenly spaced bins in a histogram of these samples by the following process:

(1) Set $H_i=0$, for $i=1,\dots,B$.
(2) For each $k=1,\dots,N_y$, find the $i=1,\dots,B$ in which $\frac{i-1}{B}\le\hat{r}_{t_k}\le\frac{i}{B}$ and set $H_i=H_i+1$.

This histogram will be refered to as the multilevel PIT histogram (MLPIT) for the remainder of this article. The MLMC approximation that derives the ensemble forecast $\{X^i_{F,t_k}\}_{i=1,\dots,N}$ can then be described as calibrated with respect to the target distribution if $H_i\approx N_y/B$ for each $i=1,\dots,B$. Thus, this can be used to test the variance and biasedness of the ensembles with respect to the target distribution. If the histogram is convex then the ensembles are said to be overdispersed, whereas if it is concave then the ensembles are said to be underdispersed and if it is skewed then there exists a bias in the ensembles (Carney and Cunningham, 2006). This is therefore a very appropriate way to clarify whether there is any additional bias from the cancellation of intermediate estimators in a MLMC approximation, thus negating the telescoping sum of expectations in (4), although this is not demonstrated in this article.

Example: The following linear mean reverting OU process, $X_t\in\mathbb{R}$,

$$\mathrm{d}X_t=\alpha(\mu-X_t)dt+\sigma^2\,\mathrm{d}W_t, \tag{19}$$

over time time interval $t\in[0,T]$, where W_t is a univariate Brownian motion, will be used alongside pre-defined scenarios of calibration for a MLMC approximation to the forecast distribution to provide a demonstration of the proposed method. We let the observations come from the above model, discretized with time step $h=2^{-5}$, with $\alpha=0.1$, $\mu=0$ and $\sigma^2=0.1$. In this example, an Euler--Maruyama numerical scheme will be used to discretize the OU process. To frame this problem in a likely forecasting setting, we first choose the fixed finest resolution that we desire, $L=4$ and so $l\in[0,4]$. A maximum computational expense that we are allowed to use on propagating the entirity of samples in each level of the ensemble hierarchy, $C_{\max}=1.536\times10^7$, is then set. The cost of each sample in the lth difference estimator is $\big\lceil Th_l^{-1}(1+1/2)\big\rceil$ (as all but the first

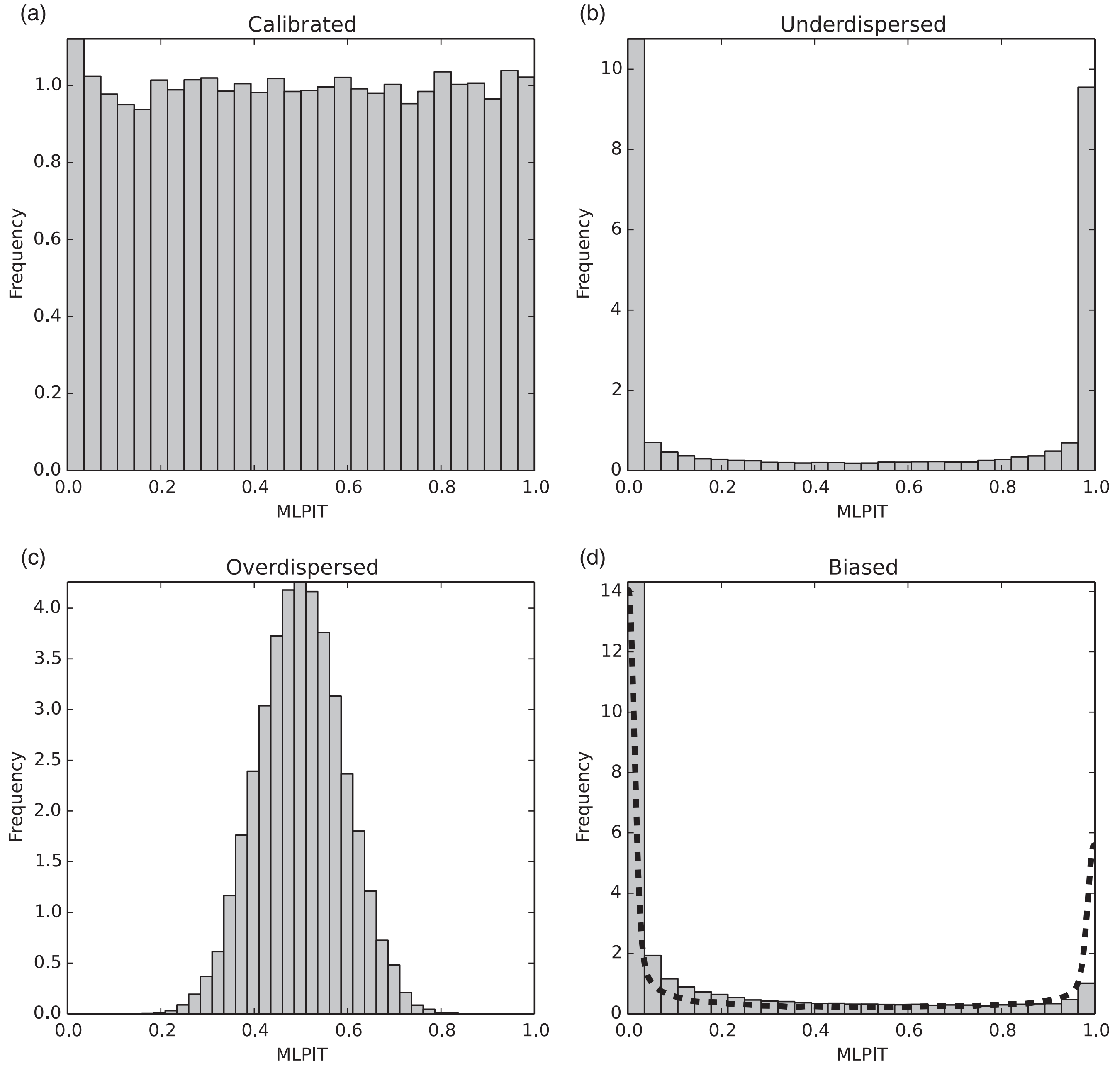

Figure 1. Multilevel probability integral transform histograms, using the ensemble $\{X_F^i\}_{i=1,\dots,N}$, of the linear OU process for the four different calibration scenarios: calibrated (a), underdispersed (b), overdispersed (c) and biased (d). The dashed line on the Biased scenario plot shows a smoothed kernel of the PIT histogram generated from the actual stationary forecast and target distributions.

difference estimators in (3) require coarse and fine time steps of the discretization), where $h_l = 2^{-1-l}$ and so N_l is given by

$$N_l = \left\lfloor \frac{C_{\max}}{T\left(h_l^{-1}\,(1+1/2)\right)} \right\rfloor = \left\lfloor \left(\frac{2}{3}\right) C_{\max} T^{-1} h_l \right\rfloor. \quad (20)$$

This corresponds to $N_0 = 2^7$. The arbitrary number of samples X_F^i to draw from the MLMC approximation to the inverse CDF is set to $N = 8N_0 = 2^{10}$.

Pairs of samples from coarse and fine ensembles in each difference estimator in (3) are positively coupled by using the same underlying Brownian motion, as in Giles (2008). The models are run over times $t \in [0, 40\,000]$ (the long run time is to give the stationary distributions a chance to be simulated) and observations are collected at $t_k = k$, $k \in [1, 40\,000]$. At each of these times, a single ensemble forecast is generated from the hierarchy of ensembles that build up the MLMC approximation to the forecast distribution and is used to verify the calibration of the approximation. Model parameters for four experimental set ups are given as follows: $\alpha = 0.1, \sigma^2 = 0.1, \mu = 0$ for the calibrated scenario, $\alpha = 0.1$, $\sigma^2 = 0.02$, $\mu = 0$ for the underdispersed scenario, $\alpha = 0.1, \sigma^2 = 0.5, \mu = 0$ for the overdispersed scenario and $\alpha = 0.4, \sigma^2 = 0.1, \mu = 0.2$ for the biased scenario.

This set up allows us to establish that the correct calibration behaviour is being shown by the MLPIT histogram for each of the scenarios; however, we will also compare this with the PIT histogram using just the finest ensemble, although this is not the primary goal of the section. Figures 1 and 2 show the MLPIT and PIT histograms, respectively, for the four scenarios of calibration listed above. Due to the small number of samples in the finest ensemble, the PIT histogram can only represent a very small number of bins of probability. Both show similar general behaviour for the cases above.

We can derive the stationary distribution to both the forecast distribution and the target distribution from the model specifications above using the Fokker--Planck equation corresponding to (19). One notes that the stationary forecast distribution using the Biased scenario model above is given by $f \sim N\left(0.2, \frac{1}{8}\right)$ and the stationary target distribution is given

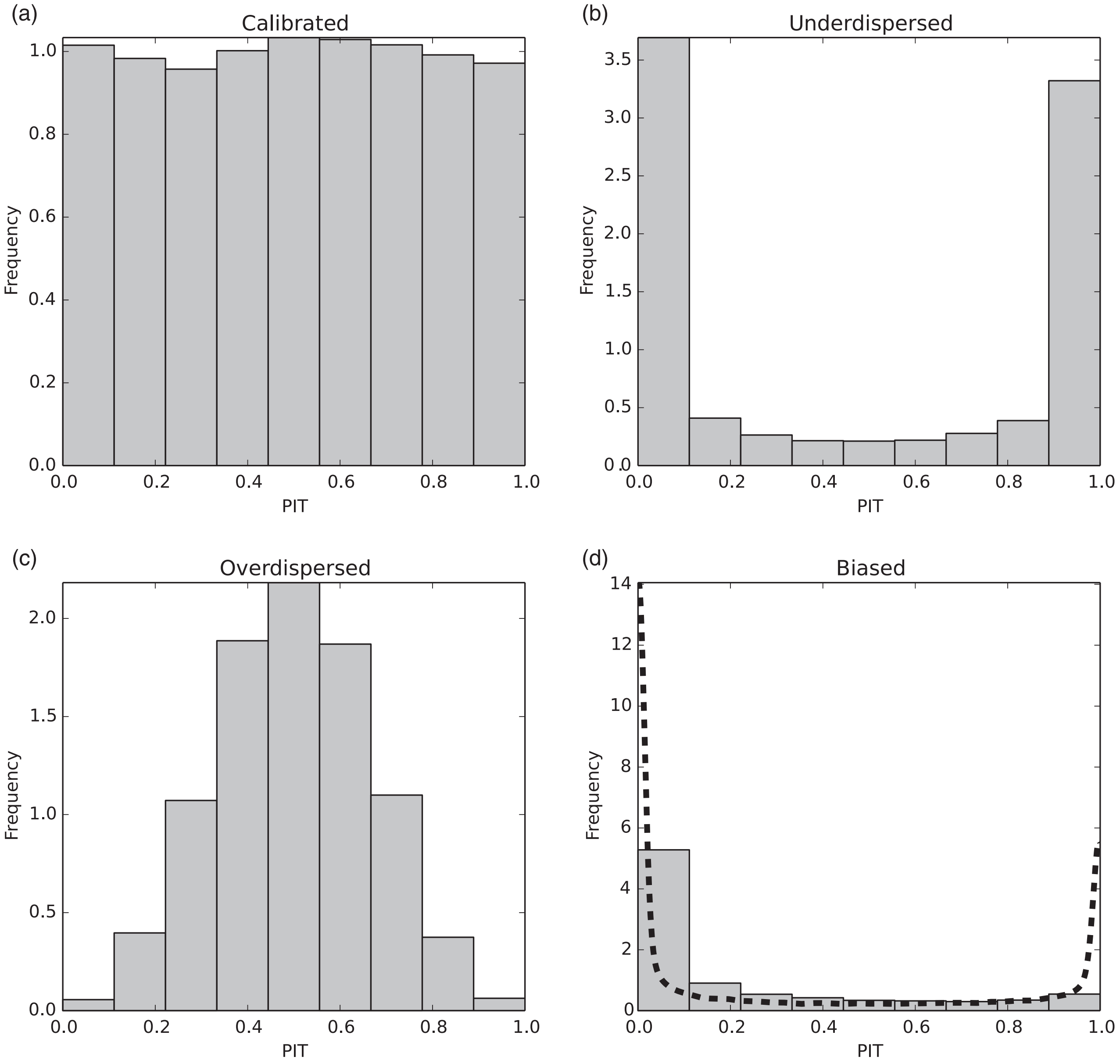

Figure 2. Probability integral transform histograms, using just the finest ensemble $\{X_L^i\}_{i=1,\ldots,N_L}$, of the linear OU process for the four different calibration scenarios: calibrated (a), underdispersed (b), overdispersed (c) and biased (d). The dashed line on the Biased scenario plot shows a smoothed kernel of the PIT histogram generated from the actual stationary forecast and target distributions.

by $y \sim N\left(0, \frac{1}{2}\right)$ (as the general form is $\sim N\left(\mu, \frac{\sigma^2}{2\alpha}\right)$). Thus the actual PIT histogram can be generated by taking an arbitrarily large number of samples of $F(y)$, where F is the CDF of f. A smoothed density kernel of this histogram is superimposed on the corresponding empirical PIT histograms for the single-level and multilevel approximations. The empirical histograms approximately match this; however, due to the lack of samples in the finest ensembles, the single-level histogram is not as clear regarding the type or magnitude of bias as shown by the MLPIT histogram. This is due to the lack of probability bins in a small, single finest ensemble $(N_L + 1)$ and one would still suffer from similar problems if using interpolation techniques in between the limited number of samples of this ensemble. The MLMC approximations of the forecast distributions and associated histograms are numerically biased (proportional to the finest time step) from this exact PIT histogram, due to the use of a numerical discretization, and so are expected to be slightly different. Despite this, one can clearly interpret the calibration and identify the extent and type of such bias in the MLMC approximations to forecast distributions with more clarity using the MLPIT histogram technique proposed here than using standard methods with the small finest ensemble.

4. Conclusion and outlook

This work has discussed the benefits of generating an ensemble forecast from Multilevel Monte Carlo (MLMC) approximations to statistics of random variables representing forecast distributions. The proposed procedure to do this is simple and easily implemented. The calibration of this ensemble forecast has also been examined. Ensemble forecasts provide a simple methodology of deriving empirical estimates to associated distribution functions. The ensemble hierarchy that forms the computationally efficient MLMC approximations to an arbitrary statistic of the forecast distribution is assumed already to have been generated, in preparation for forecasting. It is anticipated that, in real forecasting practice, this hierarchy of ensembles would simply be generated by using the maximum ensemble sizes affordable at each level of resolution. The ensemble forecast calibration verification technique takes the entire multilevel hierarchy into account when using the proposed methodology.

Calibration of this ensemble forecast is assessed using the PIT histogram after this single ensemble is generated from the ensemble hierarchy. Thus we have stated what it means for a MLMC approximation to be calibrated with respect to a target distribution. This can be used to evaluate many properties of a MLMC approximation to a forecast distribution, including biases (and their type) from intermediate terms in the MLMC telescoping sum of estimators, variances of the approximation and potentially even distribution multimodal feature detection.

Acknowledgements

Alastair Gregory was supported by the Science and Solutions to a Changing Planet DTP and the Natural Environment Research Council. He was also supported by the Mathematics of Planet Earth CDT. The authors of this article thank Chris Ferro (Exeter) for the range of helpful conversations and advice he gave in the development of this research. Thanks also go to the reviewers of this article for their constructive feedback on the manuscript.

References

Bierig C, Chernov A. 2016. Approximation of probability density functions by the Multilevel Monte Carlo Maximum Entropy method. *J. Comput. Phys.* **314**: 661–681.

Carney M, Cunningham P. 2006. ‘Evaluating density forecasting models’, *Trinity College Dublin, Dublin, Department of Computer Science*, **21**.

Cliffe KA, Giles MB, Scheichl R, Teckentrup AL. 2011. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Comput. Visual. Sci.* **140**: 3–15.

Elfverson D, Hellman F, Målqvist A. 2016. ‘A multilevel Monte Carlo method for computing failure probabilities’. *SIAM/ASA J. Uncertainty Quantification* **4**(1): 312–330.

Giles MB. 2008. Multilevel Monte Carlo path simulation. *Oper. Res.* **560**: 607–617.

Giles M. 2013. atlMultilevel Monte Carlo methods. In *Monte Carlo and Quasi-Monte Carlo Methods*, Dick J, Kuo, F, Peters G, Sloan I. (eds.) **65**: 83–103. Springer: Berlin, Heidelberg.

Giles MB. 2015. Multilevel Monte Carlo methods. *Acta Numer.* **24**: 259.

Giles MB, Nagapetyan T, Ritter K. 2015. Multilevel Monte Carlo approximation of distribution functions and densities. *SIAM/ASA J. Uncertainty Quantification* **30**: 267–295.

Gneiting T, Balabdaoui F, Raftery AE. 2007. Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)* **690**: 243–268.

Gregory A, Cotter CJ. 2016. ‘A seamless multilevel ensemble transform particle filter’. arXiv preprint arXiv:1611.00266.

Gregory A, Cotter CJ, Reich S. 2016. Multilevel ensemble transform particle filtering. *SIAM J. Sci. Comput.* **380**: A1317–A1338.

Jasra A, Kamatani K, Law KJH, Zhou Y. 2015. ‘Multilevel particle filter’. arXiv preprint arXiv:1510.04977.

Ma Y, Genton MG, Parzen E. 2011. Asymptotic properties of sample quantiles of discrete distributions. *Ann. Inst. Stat. Math.* **630**: 227–243.

Wilson D, Baker RE. 2016. ‘Multi-level methods and approximating distribution functions’. *AIP Advances*, **6**(7): 075020.