IMPERIAL

Imperial College London
Department of Mathematics

Approximate Bayesian Computation with Proper Scoring Rules: An Inference Framework for Calibrated Probabilistic Forecasts

Ashley Turner

CID: 06010243

Supervised by Dr Robin Ryder

28 August 2025

Submitted in partial fulfilment of the requirements for the MSc in Statistics at Imperial College London

The work contained in this thesis is my own work unless otherwise stated.

Signed: Ashley Turner Date: 28/08/2025

Acknowledgements

Thank you to my family and friends for their continued support and encouragement throughout my studies. I would like to thank my MSc supervisor, Dr Robin Ryder, for his invaluable guidance and support throughout this project. I am also grateful to my incoming PhD supervisors, Dr Nicola Gnecco and Dr Oliver Ratmann, for their early enthusiasm and direction. Special thanks to Martin Andrae and Joel Oskarsson (Linköping University, Sweden) for providing the pre-trained U-Net model used in this research and to Christopher Bülte (LMU Munich, Germany) for his helpful discussion on Random Field Perturbations. I would also like to acknowledge the computational resources provided by the Imperial College Research Computing Service.

Abstract

Uncertainty quantification in complex systems, where likelihood functions are intractable or computationally prohibitive, requires careful modelling decisions and precise calibration of model parameters. This report introduces Score-ABC, a likelihood-free framework utilising proper scoring rules with approximate Bayesian computation (ABC) algorithms to conduct inference on the parameters of calibrated probabilistic forecasting models. Unlike traditional ABC methods that rely on crafted summary statistics and distance measures, Score-ABC aligns the inference procedure with forecast performance criteria directly: only accepting parameter proposals which lead to forecasting distributions producing adequate scores when evaluated against historical observations.

The discussed framework is applied to calibrate a Random Field Perturbations (RFP) probabilistic weather forecasting model via three algorithmic variants; Gibbs-ABC, SMC-ABC and optimisation-based ABC tested each with two proper scoring rules; the energy score and the continuous ranked probability score. Score-ABC, when used to calibrate RFP, converges consistently and produces a significant performance boost versus uncalibrated models, resulting in competitive and spatially coherent probabilistic forecasts.

This report shows that: Score-ABC provides a theoretically-grounded inference procedure which is adaptable to different algorithmic variations, calibrated RFP procedures achieve competitive probabilistic forecasting performance, the framework can be operated efficiently despite high-dimensional forecast and parameter spaces.

1 Introduction

Central Research Question: *How can proper scoring rules act as principled discrepancy measures in approximate Bayesian computation schemes to enable likelihood-free inference of calibrated probabilistic forecasts?*

Probabilistic forecasting is an essential part of modern statistics, with particular importance in domains such as economics, meteorology, finance, epidemiology and climate science. In such scenarios, practitioners seek to produce not only point estimates of future states but also a quantification of the uncertainty associated with possible future events (Gneiting and Katzfuss, 2014).

Although many forecasting regimes begin with a powerful deterministic model, for example a set of physical equations or a machine learning system mapping present states to future outcomes, such systems are inherently limited. They do not account for sources of uncertainty such as chaotic or random underlying system dynamics, observation error or model misspecification. To address this, practitioners often resort to augmenting the base deterministic forecast with some (parameterised) stochastic component (Bülte et al., 2025). Rather than point predictions, the resulting model yields approximations of predictive probability distributions conditional on the current or past system states, but introduces the difficult challenge of using the data to learn the form of augmentation and the associated appropriate parameter choices.

This difficulty presents itself clearly in the context of model inference. In numerous complex systems, the likelihood function of the probabilistic forecasting model is unavailable. This rules out the use of both analytical solutions and standard likelihood-based inference methods such as Markov Chain Monte Carlo simulation or variational inference (Robert and Casella, 2004). In particular, forecasting systems often only define predictive distributions implicitly through sampling/simulation (for example via ensemble models (Palmer, 2002)). This issue is particularly acute in the context of operational weather forecasting (the running example used in this report) where chaotic, high-dimensional and sequential models routinely make obtaining a model likelihood impossible.

Approximate Bayesian Computation (ABC) provides a general framework to carry out inference whilst avoiding the hurdle of an intractable likelihood (Marin et al., 2012). Rather than directly evaluating the likelihood, ABC works via forward-pass simulation: a candidate parameter is proposed, synthetic data is generated under this proposal and the candidate parameter is accepted or rejected based on a discrepancy measure between the observed and simulated data. Typical ABC methods rely on tailored summary statistics and distance metrics to produce the discrepancy measure. Whilst this allows for flexible implementation, can lead to concerns over arbitrary selection of summary statistics and distance metrics and the validity of the resulting posterior approximations.

As well as the problem of intractable likelihoods, deterministic models augmented with stochastic components are subject to a deeper, structural criticism: the nature of the augmentation is frequently weakly justified and insufficiently represents the system's generative properties. In complex systems, the underlying sources of uncertainty are not identifiable, let alone suitable to justified parametric modelling. This is a particular problem in cases exhibiting chaotic or high-dimensional behaviour (such as atmospheric

dynamics). Because of this, it can be more appropriate to calibrate the empirical *behaviour* of model predictions toward desirable properties rather than aiming to derive an accurate forecasting regime from first principles. This forecast behaviour can be assessed via proper scoring rules, functions which evaluate the quality of a probabilistic forecast versus the realised outcome in such a way that honest forecasting (reporting the true underlying distribution) is optimal (Gneiting and Raftery, 2007). Proper scoring rules are widely used for forecast evaluation and are heavily researched, but their usage within likelihood-free inference methods is underexplored.

This report develops an alternative ABC framework to be applied to probabilistic forecasting models, where the discrepancy measure is defined through proper scoring rules. By using proper scoring rules to evaluate the empirical performance of forecast distributions, one can create a principled rejection criterion within the ABC algorithm, resulting in a constrained posterior parameter distribution calibrated to the proper scoring rule: candidate parameter values are accepted only if their associated forecast distributions achieve adequate score values when evaluated against historic observations.

This scoring-rule-based approach has several advantages. Firstly, it aligns the inference procedure directly with forecast performance criteria: the inference results represent adherence to the proper scoring rule which, by definition, enforces honest forecasting (Gneiting and Raftery, 2007). As well as this, it avoids a reliance on arbitrarily chosen summaries and provides a mechanism for posterior approximation which is theoretically interpretable (via score decomposition into meaningful components (Siegert, 2017)).

This report develops an ABC framework based on proper scoring rules and introduces algorithmic variants designed to address specific challenges such as high parameter dimensionality and memory-intensive workloads. These include Gibbs-like parameter proposals (Clarté et al., 2021), SMC-like adaptive proposals (Sisson et al., 2007) and wider implementation strategies for memory-efficient simulation and evaluation. The methodology is demonstrated to calibrate the parameters of a stochastic augmentation (random field perturbations (Magnusson et al., 2009)), applied to a deterministic machine learning weather prediction (MLWP) model. The deterministic U-net model used in this project is a debiased adaptation of that presented by Andrae et al. (2025), itself a reimplementation the model originally developed by Karras et al. (2022) and trained on global ERA5 data (Hersbach et al., 2020).

This report begins with an overview of preliminary topics in Section 2 including: proper scoring rules, approximate Bayesian computation, machine learning weather prediction

and uncertainty quantification techniques. Attention is given to the difficulties faced during uncertainty quantification in deterministic models and conducting inference in complex systems.

Section 3 (Methods) then covers the central Score-ABC framework, beginning with the core algorithm and subsequently discussing more sophisticated and situational variants. An application of Score-ABC to random field perturbations (RFP) for MLWP is described, including implementation details such as vectorisation and memory management techniques. Reference uncertainty quantification methods are also introduced to provide a baseline for evaluation.

Section 4 (Results) presents the empirical results of calibrating an RFP uncertainty quantification method applied to a deterministic U-Net MLWP model. The results show posterior distributions, convergence behaviour and comparative forecast performance across different scoring rules and algorithms.

Section 5 brings together the key findings of this report, acknowledges limitations of the framework and examines potential future research directions.

This work makes contributions to the intersection of probabilistic forecasting and approximate Bayesian computation by:

1. Introducing proper scoring rules as discrepancy measures in ABC procedures, aligning likelihood-free inference with forecast performance

2. Presenting computationally efficient variants of this procedure (Gibbs, SMC and optimisation-based ABC), able to handle higher-dimension parameter and forecast spaces

3. Empirically validating a Score-ABC calibrated RFP MLWP model to show competitive forecast performance while maintaining spatial coherence

4. Comparing the different algorithmic and score function variants to gain insight into their relative strengths and weaknesses

5. Highlighting how recent computational advances in machine learning weather prediction have created significant new opportunities for simulation-based methods which were previously infeasible

2 Background

Before presenting the proposed framework, we begin with a discussion of concepts foundational to the work in this report.

2.1 Proper Scoring Rules

Scoring rules provide a quantitative response to the question "How good was the predicted forecast versus what actually happened?" and are necessary to evaluate the performance of uncertainty quantification methods. A scoring rule $\mathcal{S}$ compares a probabilistic forecast $\pi_\theta(x_{t+1} \mid x_t)$ with the observed data $\mathcal{D} = \{(x_{t+1}, x_t)\}_{t=1}^{T}$ to map to a real number score $\mathcal{S}(\pi_\theta, \mathcal{D}) : \mathcal{P}(\mathcal{X}) \times \mathcal{X} \to \mathbb{R}_+$ where $\mathcal{P}(\mathcal{X})$ is a probability measure on $\mathcal{X}$.

A *proper scoring rule* is a scoring rule $\mathcal{S}$ such that the best expected score is achieved when the predicted distribution is exactly the underlying target distribution, as expressed in Equation (1).

$$\mathbb{E}_{X\sim\pi}\left[\mathcal{S}(\pi, x)\right] \leq \mathbb{E}_{X\sim\pi}\left[\mathcal{S}(\pi_\theta, x)\right] \quad \text{for all } \pi_\theta \in \mathcal{P}(\mathcal{X}),\ \pi \in \mathcal{P}(\mathcal{X}) \tag{1}$$

2.1.1 Brier Score

For a binary event with forecast probability $f_t \in [0, 1]$ and outcome $o_t \in \{0, 1\}$ at times $t = 1, \ldots, N$, the (empirical) Brier (1950) score is

$$\mathrm{BS} \;=\; \frac{1}{N}\sum_{t=1}^{N}\big(f_t - o_t\big)^2, \tag{2}$$

A key property of (2) is the Murphy (1973) decomposition

$$\mathrm{BS} \;=\; \mathrm{REL} - \mathrm{RES} + \mathrm{UNC}, \tag{3}$$

where reliability $\mathrm{REL} \geq 0$ penalises deviations between forecast probabilities and observed event frequencies within probability bins, resolution $\mathrm{RES} \geq 0$ rewards the ability to discern outcomes beyond simple long-term rates and uncertainty $\mathrm{UNC} \geq 0$ represents the marginal variance of the binary outcome.

2.1.2 Continuous Ranked Probability Score (CRPS)

Given a univariate continuous variable, its forecast cumulative distribution function F and an observed value x, the Continuous Ranked Probability Score (CRPS) is defined as

$$\mathrm{CRPS}(F, x) = \int_{-\infty}^{\infty} \left(F(y) - \mathbb{I}\{y \geq x\}\right)^2 \, dy, \tag{4}$$

where $\mathbb{I}\{y \geq x\}$ is the indicator function. Lower CRPS values indicate better calibrated (forecast probabilities align with observed event frequencies) and sharper (the ability of the forecast to assign probabilities to events beyond the long-run observed frequencies) probabilistic forecasts. It is the integral of Brier scores (i.e., the quadratic loss) over all threshold events (Hersbach, 2000).

Useful Properties and Observations

- Gneiting and Raftery (2007) provide a more intuitive form:

$$\mathrm{CRPS}(F, x) = \mathbb{E}|X - x| - \tfrac{1}{2}\,\mathbb{E}|X - X'|, \tag{5}$$

  where $X, X' \sim F$ with $\mathbb{E}\,|X| < \infty$, showing that CRPS reduces to the mean absolute error (MAE) for deterministic forecasts.

- For an ensemble forecast $\{x_1, \ldots, x_m\}$ (e.g., a sample from F), an unbiased estimator obtained by replacing expectations in (5) is

$$\widehat{\mathrm{CRPS}}_K(x_1{:}x_K; x) = \frac{1}{K}\sum_{i=1}^{K}|x_i - x| - \frac{1}{2K^2}\sum_{i=1}^{K}\sum_{j=1}^{K}|x_i - x_j|, \tag{6}$$

- Siegert (2017) provides a simple derivation of the CRPS decomposition, analogous to the classic decomposition of the Brier score (Murphy, 1973),

$$\mathrm{CRPS} = \mathrm{REL} - \mathrm{RES} + \mathrm{UNC}, \tag{7}$$

  with non-negative reliability (REL), resolution (RES), and uncertainty (UNC) terms.

- Given a v-variate forecast with marginals $F^{(j)}$ and vector observation $x = (x^{(1)}, \ldots, x^{(v)})$, the variable-wise averaged CRPS

$$\mathrm{CRPS}_{\mathrm{avg}}(F, x) \;=\; \frac{1}{v}\sum_{j=1}^{v} \mathrm{CRPS}\big(F^{(j)}, x^{(j)}\big) \tag{8}$$

  is proper for the collection of individual marginal distributions but cannot diagnose or reward cross-component dependence. Consequently, $\mathrm{CRPS}_{\mathrm{avg}}$ should not technically be used as the sole objective when joint calibration is required; multivariate proper scores such as the energy score (Gneiting and Raftery, 2007) are preferable in contexts where strict rigour is required. Despite this, the interpretability and relative fame of CRPS result in it often still being used as the standard scoring rule to evaluate probabilistic forecasts, even in multivariate settings (Bülte, 2025).

2.1.3 Energy Score

Let F be a probability distribution on $\mathbb{R}^d$ and $x \in \mathbb{R}^d$. The energy score (ES) is

$$\mathrm{ES}(F, x) \;=\; \mathbb{E}\,\|X - x\| \;-\; \tfrac{1}{2}\,\mathbb{E}\,\|X - X'\|, \tag{9}$$

with $X, X' \sim F$ i.i.d. and $\|\cdot\|$ denoting the Euclidean norm. The ES is a strictly proper scoring rule on $\mathbb{R}^d$ which coincides with the CRPS in $d = 1$ (Gneiting and Raftery, 2007), such that ES is often considered a natural extension of CRPS. The underlying metric is the energy distance, which equals zero iff the forecasted and true distributions coincide (Székely and Rizzo, 2013).

For an ensemble $\{x_1, \ldots, x_m\} \subset \mathbb{R}^d$, an estimator is

$$\widehat{\mathrm{ES}}_m(x_{1:m}; x) = \frac{1}{m}\sum_{i=1}^{m}\|x_i - x\| - \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\|x_i - x_j\|, \tag{10}$$

2.2 Machine Learning Weather Prediction (MLWP)

Machine learning weather prediction is an emerging field within the wider domain of weather and climate forecasting. Recent computational and architectural advancements have led to leaps in the abilities of data-driven models, to the point where cutting-edge MLWP models are beginning to match and even out-compete their traditional

physics-based, numerical counterparts (Bouallègue et al., 2024). One advantage of data-driven models is that forward simulation is orders of magnitude faster than in traditional numerical weather prediction (NWP) models, which rely on solving many Navier-Stokes differential equations in tandem. The computational speedup provided by competitive MLWP methods has opened the door to simulation-based inference methods previously overlooked due to their operational impossibility under NWP regimes. In particular, both graph-based and diffusion-based models have seen notable recent success. However, although some projects have seen recent success (Price et al., 2025), most efforts have thus far been directed at deterministic forecasting, with only limited consideration given to uncertainty quantification. Deterministic models are denoted in this report by $f : \mathcal{X} \to \mathcal{X}, \quad \mathcal{X} := \mathbb{R}^{d \times d \times v}$, where $d \times d$ is the dimensionality of the spatial field (arbitrarily presented as square for simplicity of notation) and v is the number of atmospheric variables in the model. The function f in this case is considered Markov and predicts only a single step forward, but the notation can be easily extended to allow for longer-range temporal dependencies and structures.

2.3 Ensemble Forecasting Models

An ensemble forecast represents predictive uncertainty by passing a collection of perturbed initial states and/or model configurations (parameters, architectures, initialisations...) through a deterministic model, generating a sample from the predictive distribution. Ensemble forecasting is standard in weather prediction, with perturbations designed to sample the fastest-growing directions of forecast error and uncertainty. Ensemble forecasts act as a Monte Carlo approximation to the target conditional predictive distribution.

2.3.1 Uncertainty Quantification in MLWP

Uncertainty quantification takes on a number of forms within the field of MLWP and is crucial for optimal decision-making in real-world uses of weather forecast data. The following section outlines some common approaches involving the augmentation of deterministic models with a parameterised stochastic component. This report is not concerned with purpose-built probabilistic models such as those based on Bayesian neural networks or probabilistic generative diffusion models.

Initial Condition (IC) Perturbations Initial condition (IC) perturbations refer to a class of uncertainty quantification methods which add a parameterised stochastic component $e_t \in \mathbb{R}^{d\times d\times v}$ to the *input* $x_t \in \mathbb{R}^{d\times d\times v}$ of a deterministic model f: $f(x_t + e_t)$. This results in the forecasting distribution $x_{t+1} \sim \pi_\theta(x_{t+1} \mid x_t)$. Generic choices include: naive Gaussian white noise, spatially correlated Gaussian fields (obtained by filtering white noise through stationary covariance kernels) and data-assimilation-derived perturbations (such as random field perturbations (Magnusson et al., 2009)). The resulting ensemble $\{f(x_t^{(k)})\}_{k=1}^K$ serves as a Monte Carlo approximation of $\pi_\theta(\cdot \mid x_t)$. The construction and implementation of $e_t^{(k)}$ critically affect forecast properties such as calibration, spread and reliability, motivating principled parameter inference for perturbation tuning (Bülte et al., 2025).

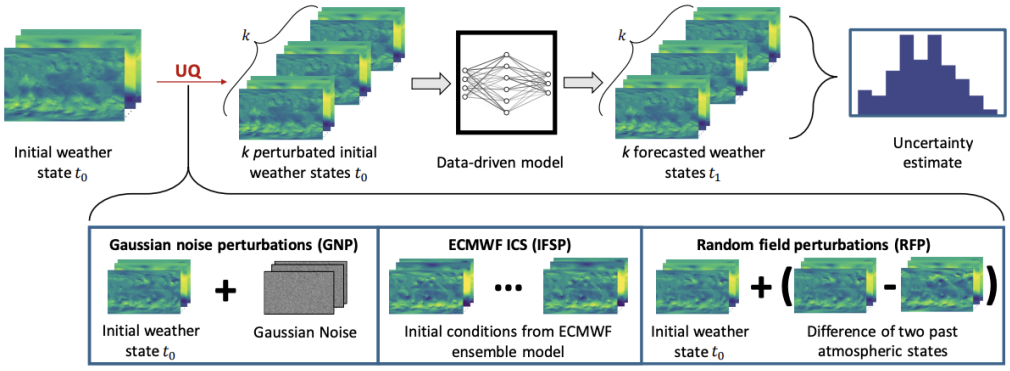

Figure 1: Bülte et al. (2025)'s illustration of several IC perturbation methods

Random Field IC Perturbations (RFP) Random field perturbations aim to preserve physical balances by sampling perturbations from historical atmospheric states rather than injecting completely synthetic noise. Denoting the past field of variable j at time τ as $z^{(j)}(\tau) \in \mathbb{R}^{d\times d}$ and defining differences $\Delta z^{(j)}(\tau_1, \tau_2) = z^{(j)}(\tau_1) - z^{(j)}(\tau_2)$ for independently drawn $\tau_1 \neq \tau_2$ from a seasonally and diurnally matched pool. With an energy[1] norm $\|\cdot\|_E$ and scale parameter $\alpha_j > 0$, construct

$$e_t^{(k,j)} = \alpha_j \frac{\Delta z^{(j)}(\tau_1^{(k)}, \tau_2^{(k)})}{\left\|\Delta z^{(j)}(\tau_1^{(k)}, \tau_2^{(k)})\right\|_E}, \quad j = 1, \ldots, v, \tag{11}$$

[1] In this report's implementation, the energy norm has been (crudely) approximated by the Euclidean norm, as calculating the true energy norm per Magnusson et al. (2009) requires atmospheric variables not included in this work's dataset.

Stacking $e_t^{(k)} = (e_t^{(k,1)}, \ldots, e_t^{(k,v)})$ results in spatially coherent and physically plausible IC perturbations (Magnusson et al., 2009).

Choosing the value of $\alpha = (\alpha_1, \ldots, \alpha_v)$ is a critical yet challenging step in the creation of a well-calibrated uncertainty quantification method. This report specifically focuses Score-ABC as a likelihood-free inference procedure to examine α for the purpose of model calibration.

2.4 Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) refers to a class of likelihood-free methods for approximating the posterior distribution of models with an unavailable likelihood (for example because it is unknown, intractable or prohibitively expensive to evaluate), but where forward simulations through the model are trivial.

$$p(\theta \mid \mathcal{D}) = \frac{p(\theta)\overbrace{\pi(\mathcal{D} \mid \theta)}^{\text{intractable}}}{\pi(\mathcal{D})} \tag{12}$$

Notation: p and π are used in tandem to distinguish between parameter-generating and data-generating distributions respectively.

ABC circumvents the issue of an unavailable likelihood, producing a direct approximation of the posterior distribution via:

- Proposing a parameter $\tilde{\theta}$
- Simulating synthetic data $\tilde{\mathcal{D}}$ under the proposed parameter $\tilde{\theta}$
- Using the synthetic data $\tilde{\mathcal{D}}$ and the observed data $\mathcal{D}$ to compute a discrepancy $\rho(\mathcal{D}, \tilde{\mathcal{D}})$
- Accepting/rejecting $\tilde{\theta}$ based on some discrepancy threshold ε

This process is formalised in Algorithm 1.

Algorithm 1 Vanilla Approximate Bayesian Computation (ABC)

Require: Prior $p(\theta)$; data-generating mechanism π_θ; observed data $\mathcal{D}$; summary statistic $s(\cdot)$; discrepancy metric $\rho(\cdot,\cdot)$; acceptance threshold $\varepsilon > 0$; desired sample size $N \in \mathbb{N}$

Ensure: Accepted parameter set $\{\theta^{(i)}\}_{i=1}^{N}$ approximating $\mathbb{P}(\theta \mid \rho(s(\tilde{\mathcal{D}}), s(\mathcal{D})) \leq \varepsilon)$

Set $i \leftarrow 0$
while $i < N$ **do**
    Sample $\tilde{\theta} \sim p(\theta)$
    Simulate synthetic data $\tilde{\mathcal{D}}$ via $\tilde{\pi}_{\tilde{\theta}}$
    Compute discrepancy $\rho\left(s(\tilde{\mathcal{D}}),\, s(\mathcal{D})\right)$
    if $\rho \leq \varepsilon$ **then**
        Save $\theta^{(i)} \leftarrow \tilde{\theta}$
        Continue
    end if
end while

In order to use a standard ABC scheme, one must establish appropriate proposal distributions, an acceptance threshold, summary statistic(s) and a discrepancy. This is typically not a straightforward process and is a key criticism of the ABC framework.

3 Methods

This report presents Score-ABC with an application to calibrating RFP parameters for probabilistic forecasting using a MLWP model, serving as a concrete example of a challenging situation in which Score-ABC can be successfully employed: a deterministic U-Net model (Andrae et al., 2025) produces quality forecasts but requires a stochastic augmentation to quantify uncertainty. The key question in this scenario is how to choose the RFP scale parameter $\alpha = (\alpha_1, \ldots, \alpha_v)$. This is challenging because the likelihood of the augmented MLWP is intractable and the nature of the interaction between the RFP augmentation and forecasting skill is unknown. Score-ABC addresses these issues by using proper scoring rules as the discrepancy metric, aligning the parameter inference procedure with the ultimate goal of producing calibrated probabilistic forecasts.

3.1 Mathematical Framework

This report considers a class of models where deterministic forecasts $f(x_t) = \tilde{x}_{t+1}$ are augmented with stochastic perturbations to provide predictive distributions. These pre-

dictive distributions are denoted as

$$\pi_\theta(x_{t+1} \mid x_t), \tag{13}$$

where $\theta \in \Theta \subseteq \mathbb{R}^p$ are the parameters of the noise model augmenting the deterministic simulator.

Given forecast-observation pairs constructed from data $\mathcal{D} = \{(x_t, x_{t+1})\}_{t=0}^{T-1}$, a proper scoring rule can be defined,

$$\mathcal{S} : \mathcal{P}(\mathcal{X}) \times \mathcal{X} \to \mathbb{R}_+ \tag{14}$$

and candidate parameters are evaluated with the time-aggregated empirical forecast score

$$\bar{\mathcal{S}}(\pi_\theta, \mathcal{D}) := \frac{1}{T} \sum_{t=0}^{T-1} \hat{\mathcal{S}}\left(\pi_\theta(\cdot \mid x_t), x_{t+1}\right). \tag{15}$$

3.1.1 Basic Algorithm: ABC with Proper Scoring Rules

In the scenario that π_θ is only available via forward simulation, the likelihood $L(\theta) = \prod_t \pi_\theta(x_{t+1} \mid x_t)$ is unavailable. This means that we can utilise approximate Bayesian computation (ABC) to conduct posterior inference (as described in Section 3.1.1). The parameter vector θ is treated as a latent variable with associated prior $\pi(\theta)$.

Given a tolerance $\varepsilon > 0$, the following framework defines a *score-informed* posterior:

$$p_\varepsilon(\theta \mid \mathcal{D}) \;\propto\; p(\theta)\, K_\varepsilon\big(\bar{\mathcal{S}}(\theta)\big), \tag{16}$$

where $K_\varepsilon : \mathbb{R}_+ \to [0, 1]$ is any nonincreasing acceptance kernel that assigns higher weight to parameters producing better scores and $\bar{\mathcal{S}}(\theta)$ is the empirical value of a proper scoring rule computed by comparing forecasts generated under θ to the observations $\mathcal{D}$.

In this rejection-based ABC algorithm, parameters are proposed from a distribution $q(\theta)$, forward simulations provide a sample under π_θ, and the proposal is accepted with probability

$$K_\varepsilon\big(\bar{\mathcal{S}}(\theta)\big)\,, \tag{17}$$

Algorithm 2 Score-Informed Approximate Bayesian Computation

Require: Proposal distribution $q(\theta)$; prior $p(\theta)$; forward simulator π_θ; observed data $\mathcal{D}$; proper scoring rule $\mathcal{S}$; tolerance $\varepsilon > 0$; acceptance kernel K_ε; desired sample size $N \in \mathbb{N}$

Ensure: Accepted parameter set $\{\theta^{(i)}\}_{i=1}^{N}$ approximating $p_\varepsilon(\theta \mid \mathcal{D}) \propto p(\theta) K_\varepsilon(\bar{\mathcal{S}}(\theta))$

Set $i \leftarrow 0$
while $i < N$ **do**
  Draw $\tilde{\theta} \sim q(\theta)$
  Generate synthetic probabilistic forecasts $\{\tilde{x}_{t+1}\}_{t=0}^{T-1} \sim \pi_{\tilde{\theta}}(\cdot \mid x_t)$
  Compute the empirical score

$$\bar{\mathcal{S}}(\tilde{\theta}) \leftarrow \frac{1}{T} \sum_{t=0}^{T-1} \hat{\mathcal{S}}\big(\pi_{\tilde{\theta}}(\cdot \mid x_t),\, x_{t+1}\big) \tag{18}$$

  Accept $\tilde{\theta}$ with probability $K_\varepsilon\Big(\bar{\mathcal{S}}(\tilde{\theta})\Big)$
  if accepted **then**
    Set weight $w^{(i)} \leftarrow \frac{\pi(\tilde{\theta})}{q(\tilde{\theta})}$
    Store $(\theta^{(i)}, w^{(i)}) \leftarrow (\tilde{\theta}, w^{(i)})$
  end if
end while

Motivation: An important property of the Score-ABC framework is immediately clear when considering the sufficiency of scoring rules as parameter summary statistics: a proper scoring rule is unlikely to be a truly sufficient statistic for a complex system. This report argues that this is not a problem and is, in fact, a false goal in an operational probabilistic forecasting context.

Using a score function as the discrepancy metric in an ABC algorithm naturally falls short in the traditional sense of the aim of an ABC summary statistic: to be sufficient for the posterior parameter. However, it is important to be clear that the Score-ABC scheme does not target the *true* parameter posterior, rather that is concerned with the *score-informed* posterior: the density over parameters which have adequate score-encoded empirical properties (e.g., calibration). Because of this, in the described probabilistic forecasting context, the lack of a sufficient statistic is not a flaw.

In the discussed modelling setting: augmenting a deterministic model with a (somewhat arbitrary) stochastic component likely results in some degree of model misspecification. Not only are such models often inadequately designed for probabilistic forecasting tasks, a 'true' and accurately specified model can be almost impossible to elucidate. Consider the case of probabilistic forecasting in a high-dimensional chaotic system: attempting to articulate the *true* data-generating mechanism is futile. When forecasting in an opera-

tional setting, targeting the true posterior of parameters under an inherently misspecified model is a false objective, motivating the abandonment of the sufficiency criterion when choosing a discrepancy metric. This project proposes employing a class of metric with concrete practical implications: proper scoring rules.

3.2 Application to Random Field Perturbations (RFP)

When using RFP to create spatially-coherent ensembles for the purposes of uncertainty quantification, the per-atmospheric-variable scale parameter α must be specified. This is the parameter of interest for this report's implementation of Score-ABC. Formally, an approximation of $p_\varepsilon(\theta \mid \mathcal{D}) \propto p(\theta)\, K_\varepsilon\big(\bar{\mathcal{S}}(\theta)\big)$ is required, where $\theta = (\alpha_1, \ldots, \alpha_v)$: the vector of per-atmospheric-variable RFP scale parameters.

The Score-ABC algorithm is run with the data-generating mechanism π_θ being an ensemble of random-field-perturbed initial states passed through a deterministic MLWP model[2]. The basic Score-ABC algorithm is outlined in Algorithm 2.

3.2.1 Challenges

Parameter Dimensionality The vector of parameters $\theta = \alpha = (\alpha_1, \ldots, \alpha_v)$ specifies perturbation magnitudes per forecast variable. In realistic applications $v \gg 1$, and naive ABC rejection falls victim to the curse of dimensionality: parameter proposals are significantly more difficult to accept as parameter dimension grows.

To counteract this, several techniques can be adapted for ABC purposes from the realm of more well-known MCMC methods. Examples of such procedures are: ABC with Gibbs-like steps and SMC-ABC.

Resource Constraints In order to compute the empirical score for a given $\tilde{\theta}$ proposal, $T-1$ time steps must be individually evaluated. Each one of these forward passes of π_θ involves simulating an ensemble of K members, through a neural network. Overall this is $K \times (T-1)$ simulations per proposal $\tilde{\theta}$. Even at conservative sample and ensemble sizes, this could involve thousands of model forward passes for a *single* parameter proposal.

[2]Many thanks to Martin Andrae (Linköping University) for providing the pre-trained U-net deterministic global forecasting model which was used in this report's implementation.

Without both effective algorithmic and implementation optimisations the cost of this simulation can be prohibitive.

Some measures to deal with the heavy computational burden are: algorithmic optimisations to limit the number of necessary proposals such as ABC with Gibbs-like steps, batch vectorisation (if memory permits) and temporal resampling (using a different subset of time steps T for each proposal acting as an unbiased estimator of the population score).

3.2.2 Algorithm: Gibbs-ABC with Proper Scoring Rules

Clarté et al. (2021) provide a solution to conduct ABC procedures in settings with significant parameter dimensionality. They propose using Gibbs-like steps (Robert and Casella, 2004) to run component-wise ABC aimed at sequentially estimating conditional parameter distributions (see Algorithm 3). The benefits of using this method are twofold: In standard ABC settings, the dimensionality of any sufficient summary statistic (used in the discrepancy metric calculation step of the ABC algorithm) is lower-bounded by the dimensionality of the parameter itself (Fearnhead and Prangle, 2012). Targeting a lower-dimensional conditional distribution within each Gibbs step allows for the summary statistic to also be of a lower dimension, meaning the impact of the curse of dimensionality on the summary statistic is reduced. Secondly, using conditional Gibbs steps also allows for parameter proposals to account for the current values of (conditionally-fixed) other components. This means that it is possible to mitigate inefficiently simulating directly from the prior: computational resources are directed more productively at higher-mass areas of the posterior by proposing from a more informative conditional prior.

Given the Score-ABC setting, it is the latter of these two properties which is of particular relevance. The first property, the dimensionality of a sufficient summary statistic, is not relevant to Score-ABC as summary statistic sufficiency is knowingly sacrificed when opting to use a score function as the discrepancy measure (as discussed in Section 3.1.1).

Let $\theta = \alpha = (\alpha_1, \ldots, \alpha_v)$. For each coordinate $i \in \{1, \ldots, v\}$ define the block-wise update that conditions on the complement θ_{-i}. Following (Clarté et al., 2021), the joint ABC rejection step is replaced with sequential one-dimensional (or possibly low-dimensional), conditionally independent accept-reject updates. At step i, propose α_i according to a conditional prior $p(\alpha_i \mid \theta_{-i})$, construct the full candidate $\theta^\star = \theta_{-i} \cup \{\alpha_i^\star\}$,

simulate T forecasts under $\pi_{\theta^\star}$, and compute the empirical score $\bar{\mathcal{S}}(\theta^\star)$ as in (15). The candidate is then accepted with probability $K_{\varepsilon_i}\big(\bar{\mathcal{S}}(\theta^\star)\big)$, where the tolerance ε_i can be coordinate-specific. Iterating these conditional steps results in a Markov chain on Θ whose stationary distribution is a score-informed Gibbs-ABC approximation to (16), while concentrating computation on high-mass regions through conditionally informative proposals.

Algorithm 3 Gibbs-like Score-ABC (component-wise ABC with proper scoring rules)

Require: Prior $p(\theta) = \prod_{i=1}^{v} p(\alpha_i \mid \theta_{-i})$; conditional priors $\{p_i(\cdot \mid \cdot)\}_{i=1}^{v}$; simulator π_θ; data $\mathcal{D} = \{(x_t, x_{t+1})\}_{t=0}^{T-1}$; proper scoring rule $\mathcal{S}$; coordinate tolerances $\{\varepsilon_i > 0\}_{i=1}^{v}$; acceptance kernels $\{K_{\varepsilon_i}\}_{i=1}^{v}$; number of sweeps $N \in \mathbb{N}$; initial $\theta^{(0)} \in \Theta$

Ensure: Markov chain $\{\theta^{(n)}\}_{n=1}^{N}$ targeting a score-informed Gibbs-ABC approximation to $p_\varepsilon(\theta \mid \mathcal{D}) \propto p(\theta)\, K_\varepsilon\big(\bar{\mathcal{S}}(\theta)\big)$

for $n \leftarrow 1, \ldots, N$ **do**
  Set $\theta^{(n,0)} \leftarrow \theta^{(n-1)}$
  for $i \leftarrow 1, \ldots, v$ **do**
    Fix $\theta_{-i}^{(n,i-1)}$ and define the conditional state space for α_i
    repeat
      Draw $\alpha_i^\star \sim p_i(\cdot \mid \theta_{-i}^{(n,i-1)})$
      Form $\theta^\star \leftarrow \theta_{-i}^{(n,i-1)} \cup \{\alpha_i^\star\}$
      Generate synthetic forecasts $\{\tilde{x}_{t+1}\}_{t=0}^{T-1}, \quad \tilde{x}_{t+1} \sim \pi_{\theta^\star}(\cdot \mid x_t)$
      Compute empirical score

$$\bar{\mathcal{S}}(\theta^\star) \;\leftarrow\; \frac{1}{T} \sum_{t=0}^{T-1} \hat{\mathcal{S}}(\pi_{\theta^\star}(\cdot \mid x_t),\, x_{t+1}) \tag{19}$$

      Accept $\alpha_i^\star$ with probability $K_{\varepsilon_i}\big(\bar{\mathcal{S}}(\theta^\star)\big)$
    until accepted
    Set $\theta^{(n,i)} \leftarrow \theta_{-i}^{(n,i-1)} \cup \{\alpha_i^\star\}$
  end for
  Set $\theta^{(n)} \leftarrow \theta^{(n,v)}$
end for

Reference Table Approach. In the original implementation of Clarté et al. (2021), the ABC-Gibbs procedure (Algorithm 3 in their paper) draws each component θ_j conditionally on the remaining parameters by simulating from its conditional prior, generating a small *reference table* of simulated datasets and selecting the parameter with the smallest discrepancy measure. This method explores the local conditional parameter space efficiently and deterministically, retaining the candidate with the best discrepancy measure and is the approach taken in this report's implementation. The reference table method does have implications for the asymptotic parameter posterior as the argmin choice ker-

nel distorts accepted parameters toward the posterior mode. Clarté et al. (2021) show that whilst the intermediate posterior generated through this procedure is not the same as the true Vanilla ABC posterior, the ABC-Gibbs algorithm with the reference table approach does converge (under some assumptions) to the true target posterior.

3.2.3 Algorithm: SMC-ABC with Proper Scoring Rules

An alternate approach to mitigating the curse of dimensionality in ABC schemes is an SMC-based algorithm. Sisson et al. (2007) describe a procedure involving the propagation of a number of particles proposed drawn from a (potentially non-prior) proposal distribution which are subsequently weighted, normalised and resampled to produce an approximation of the posterior distribution (see Algorithm 4). Let $\mathcal{S}$ be a proper score and $\bar{\mathcal{S}}(\theta)$ its empirical average (15). As before, $K_\varepsilon : \mathbb{R}_+ \to [0,1]$ is a nonincreasing kernel. SMC-ABC uses weighted particles $\{(\theta_i^{(t)}, w_i^{(t)})\}_{i=1}^M$ to construct a sequence of intermediate target distributions

$$\pi_t(\theta) \;\propto\; p(\theta)\, K_{\varepsilon_t}\big(\bar{\mathcal{S}}(\theta)\big) \tag{20}$$

Algorithm 4 SMC-ABC with proper scoring rules

Require: Prior $p(\theta)$; simulator π_θ; data $\mathcal{D}$; proper score $\mathcal{S}$; tolerances $\{\varepsilon_t\}_{t=1}^T$; acceptance kernels K_{ε_t}; proposal kernels $\{q_t(\cdot \mid \cdot)\}$; particles M

Ensure: Weighted particles $\{(\theta_i^{(t)}, w_i^{(t)})\}_{i=1}^M$ approximating $f_t(\theta) \propto p(\theta)\, K_{\varepsilon_t}(\bar{\mathcal{S}}(\theta))$

for $t = 1, \ldots, T$ **do**
  for $i = 1, \ldots, M$ **do**
    if $t = 1$ **then**
      Draw $\theta \sim p(\theta)$; forward simulate under θ to compute $\bar{\mathcal{S}}(\theta)$
      Set $\theta_i^{(1)} \leftarrow \theta$, $\bar{w}_i^{(1)} \leftarrow 1/M$
    else
      Draw ancestor $J \sim \mathrm{Cat}\big(w_1^{(t-1)}, \ldots, w_M^{(t-1)}\big)$
      Propose $\tilde{\theta} \sim q_t(\cdot \mid \theta_J^{(t-1)})$; forward simulate to compute $\bar{\mathcal{S}}(\tilde{\theta})$
      Set $\theta_i^{(t)} \leftarrow \tilde{\theta}$
      Set $\bar{w}_i^{(t)} \leftarrow p(\tilde{\theta}) / \left(\sum_{j=1}^M w_j^{(t-1)}\, q_t(\tilde{\theta} \mid \theta_j^{(t-1)})\right)$
    end if
  end for
  Normalise $w_i^{(t)} \leftarrow \bar{w}_i^{(t)} \big/ \sum_{j=1}^M \bar{w}_j^{(t)}$
end for

An adaptive proposal was used for q during implementation, drawing from a normal distribution centred at the previous step's parameter and with a scale set by a prescribed

adaptation schedule: $\tilde{\theta} \sim \mathcal{N}\big(\theta_{t-1}, \sigma_t\big)$. As well as this, ε_t was set to as an empirical quantile of the proposals.

3.3 Implementation and Practical Concerns

3.3.1 Temporal Resampling

Instead of evaluating the time-aggregated forecast score $\bar{\mathcal{S}}(\pi_\theta, \mathcal{D})$ from Equation 15 over the full horizon T spanning all available training data, one may use an unbiased estimator of $\bar{\mathcal{S}}$ by using a reduced horizon T at each evaluation, with the specific included time indices re-sampled independently at every invocation.

This means that over time, the algorithm is exposed to a large and representative sample of temporal slices without having to tackle the immense computational workload of computing forecasts over the entire set of training states.

3.3.2 Gibbs Optimiser

A pseudo-Gibbs algorithm was implemented as a comparative hybrid between the Gibbs and SMC methods. The Gibbs optimiser algorithm is identical to the Gibbs-ABC algorithm, apart from the proposal distribution, which is Gaussian with a shrinking proposal variance (as used in the SMC method). This adaptive proposal with no subsequent importance weighting means that the resulting posterior is not the same as that targeted by the SMC method, and the algorithm is more akin to an optimisation procedure. One of the consequences of this is that the Gibbs optimiser liable to become "stuck" in local minima and produce degenerate posteriors.

3.3.3 Computational Efficiency and Memory Management

Vectorisation and Memory Management To produce a single ensemble forecast π_θ, K forward passes need to be made through the deterministic forecasting model. To evaluate a single variable within a Gibbs-like step, $K \times M$ forward passes need to be made. These simulations are independent and need not happen sequentially: it is feasible to parallelise or vectorise this process, both in terms of proposals and ensemble-members. Thanks to the efficiencies of GPU architectures, vectorisation is the most performance-enhancing

method: stacks of proposals of entire ensembles can be evaluated with a single forward pass through the forecasting model.

Given the large dimension of weather models, extreme vectorisation can easily overload memory. As part of the implementation, adaptive batch sizes and differing degrees of vectorisation were employed in order to mitigate the risk of encountering crashes mid-run. Naturally, checkpointing, memory monitoring and extensive logging were all an essential part of the implementation.

3.4 Reference Methods

To evaluate the effectiveness of Score-ABC calibrated RFP forecasts, several reference methods were implemented to provide comparative baselines for the evaluation in Section 4. These methods range from naive forecasting approaches to more sophisticated probabilistic models, establishing a range of reference benchmarks for the variants of Score-ABC to be compared against.

3.4.1 Persistence

Persistence forecasting is the simplest forecasting baseline presented, predicting that future atmospheric conditions will remain identical to the most recent observation. Formally, given the current atmospheric state $x_t \in \mathbb{R}^{d\times d\times v}$, the persistence forecast is

$$\hat{x}_{t+1} = x_t. \tag{21}$$

This deterministic method provides a naive lower bound for forecast skill and serves as the reference against which improvements from more complex methods can be measured. In meteorological contexts, persistence forecasting typically performs reasonably well for very short forecast horizons (Mittermaier, 2008) but deteriorates rapidly as lead time increases due to the chaotic nature of the atmosphere.

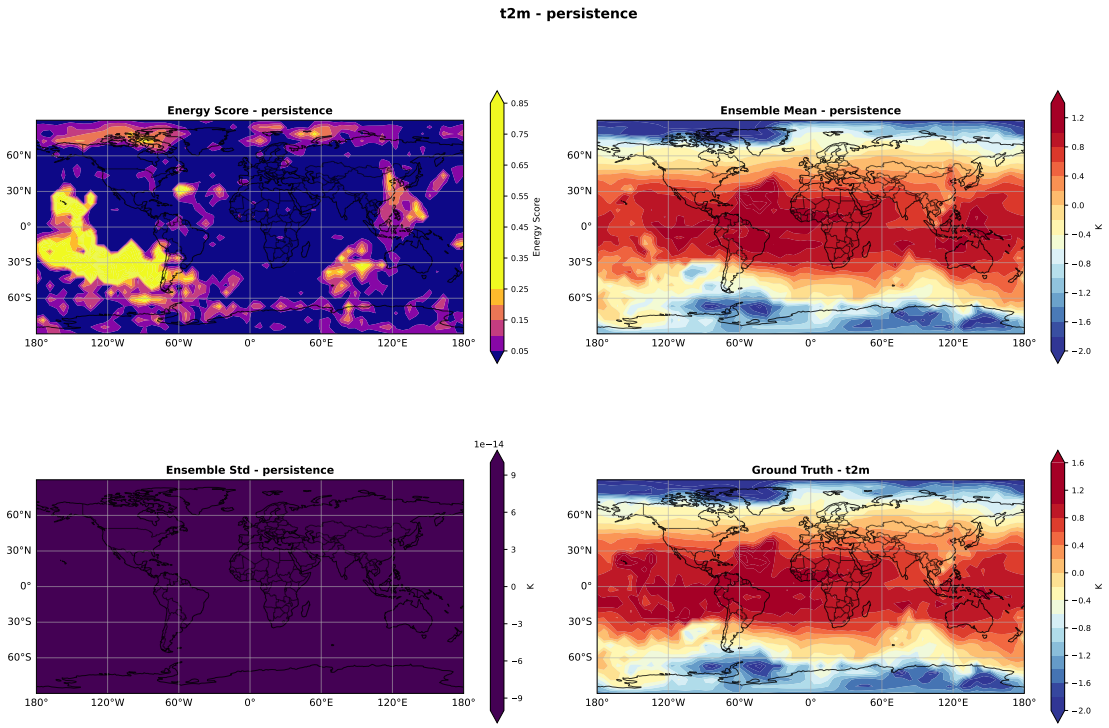

Figure 2: Example persistence forecast

3.4.2 Climatology

Climatological forecasting generates probabilistic predictions by sampling from the long-term marginal distribution of atmospheric states, ignoring current conditions entirely. The climatological mean $\mu_{\text{clim}} \in \mathbb{R}^{v \times d \times d}$ and standard deviation $\sigma_{\text{clim}} \in \mathbb{R}^{v \times d \times d}$ are computed from historical observations $\{x_\tau\}_{\tau=1}^{T}$ as

$$\mu_{\text{clim}} = \frac{1}{T} \sum_{\tau=1}^{T} x_\tau, \tag{22a}$$

$$\sigma_{\text{clim}} = \sqrt{\frac{1}{T-1} \sum_{\tau=1}^{T} (x_\tau - \mu_{\text{clim}})^2}. \tag{22b}$$

Ensemble forecasts are then generated from independent normal distributions at each grid point:

$$x_{t+1}^{(k)} \sim \mathcal{N}(\mu_{\text{clim}}, \sigma_{\text{clim}}^2), \quad k = 1, \ldots, K. \tag{23}$$

While climatological forecasts ignore all current information, they provide a baseline for assessing whether a forecasting system can extract predictive information from current states (Siegert, 2017). They also establish an upper bound on forecast uncertainty that should be improved upon by any reasonable forecasting method.

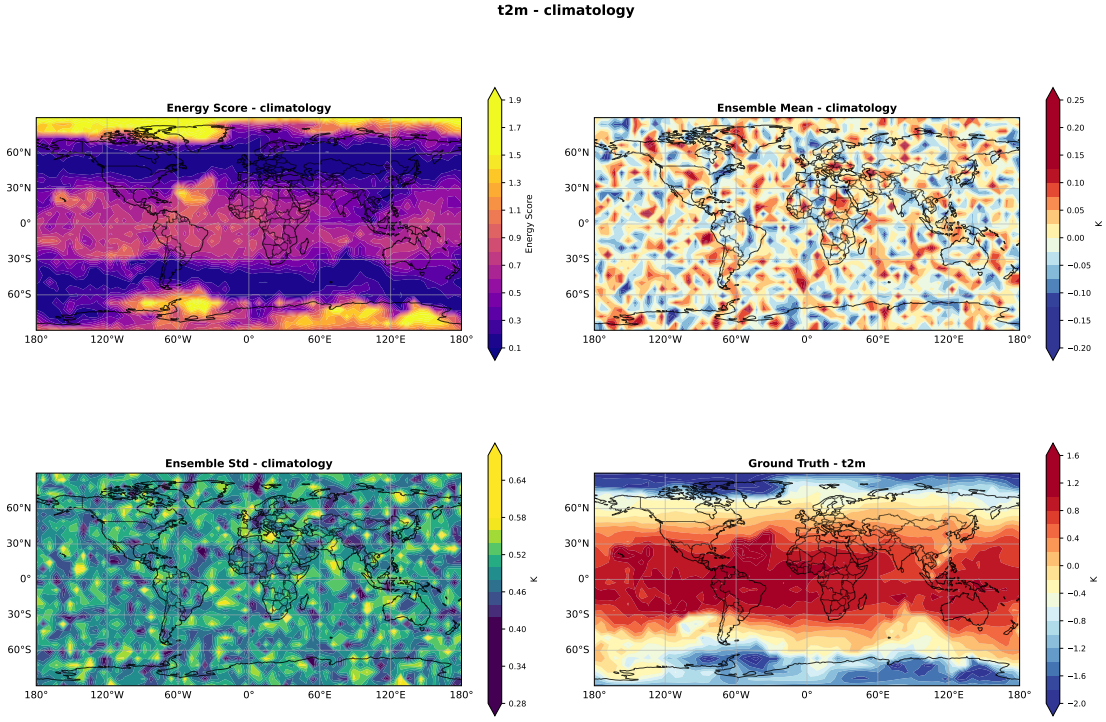

Figure 3: Example climatology ensemble forecast

3.4.3 Deterministic + Independent Gaussian Noise

This approach, suggested by Lorenz et al. (2009), augments the deterministic model f with spatially independent calibrated Gaussian noise to approximate forecast uncertainty. The method first generates a deterministic prediction $\tilde{x}_{t+1} = f(x_t)$, then adds independent Gaussian perturbations with parameters estimated via examining the empirical residuals of the deterministic forecasting model.

Empirical standard deviations $\sigma_{\text{emp}} \in \mathbb{R}^{v \times d \times d}$ are computed from model residuals on a validation set:

$$\sigma_{\text{emp}} = \sqrt{\text{Var}(f(x_t) - x_{t+1})}, \tag{24}$$

where the variance is computed element-wise across the validation samples.

Probabilistic ensemble forecasts are then constructed as:

$$x_{t+1}^{(k)} = f(x_t) + \epsilon^{(k)}, \quad \epsilon^{(k)} \sim \mathcal{N}(0, \sigma_{\text{emp}}^2), \quad k = 1, \ldots, K. \tag{25}$$

This method accurately captures the marginal uncertainty associated with each variable at each grid point but suffers from the limitation that perturbations are spatially uncorrelated, violating physical consistency and producing incoherent forecasts. Despite the resulting forecasts being unfit for operational purposes requiring spatial coherence, this method is almost perfectly marginally calibrated at every individual grid-point and serves as a useful benchmark for comparing uncertainty quantification methods.

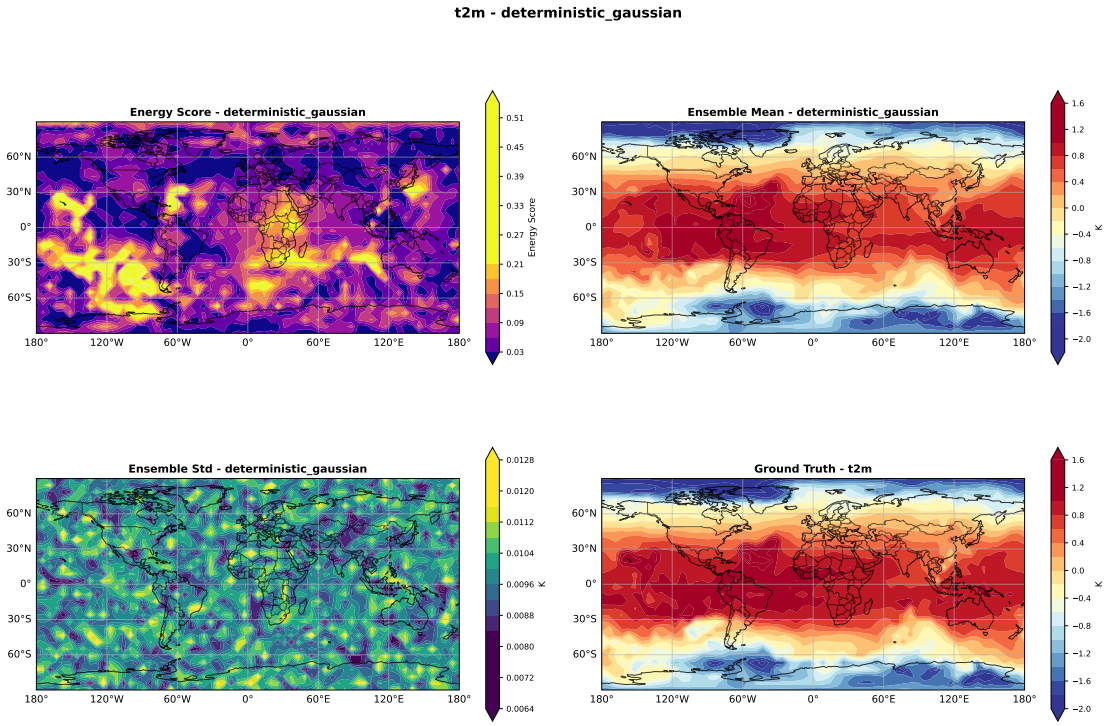

Figure 4: Example deterministic + calibrated independent Gaussian noise ensemble forecast

3.4.4 Raw RFP ($\alpha = 1.0$)

Raw RFP serves as an uncalibrated baseline for the random field perturbation method, using identical scaling parameters $\alpha_j = 1.0$ for all atmospheric variables $j = 1, \ldots, v$. This approach applies the RFP methodology described in equation (11) without any

parameter optimisation:

$$e_t^{(k,j)} = 1.0 \cdot \frac{\Delta z^{(j)}(\tau_1^{(k)}, \tau_2^{(k)})}{\|\Delta z^{(j)}(\tau_1^{(k)}, \tau_2^{(k)})\|_E}. \tag{26}$$

This method demonstrates the importance of variable-specific calibration by providing a baseline against which the benefits of Score-ABC parameter inference can be measured.

3.4.5 Prior RFP

Prior RFP represents a second uncalibrated baseline where scaling parameters are drawn randomly from the prior distribution used in the ABC algorithm. This approach provides a stochastic baseline that accounts for the expected range of scaling parameters under the prior, but without incorporating any information from the observed data. By comparing against this method, the effectiveness of Score-ABC in learning appropriate parameter values from forecast performance can be assessed. The Prior RFP method serves as an intermediate benchmark between the Raw RFP approach and fully calibrated Score-ABC methods.

4 Results

4.1 Experimental Design and Implementation

Six distinct experiments were conducted to evaluate the performance of the proposed Score-ABC framework:

1. Score-ABC with Gibbs-like steps with an energy score objective
2. Score-ABC with Gibbs-like steps with a mean CRPS objective
3. Score-ABC with SMC steps with an energy score objective
4. Score-ABC with SMC steps with a mean CRPS objective
5. Score-ABC with Gibbs optimiser steps with an energy score objective
6. Score-ABC with Gibbs optimiser steps with a mean CRPS objective

All experiments shared identical hyperparameters (except for B) as detailed in Table 1:

- $V = 5$ atmospheric variables:
  - z500 = geopotential height at 500 hPa
  - t850 = air temperature at 850 hPa
  - t2m = near-surface air temperature at 2 m above ground
  - u10 = zonal wind component at 10 m above ground
  - v10 = meridional wind component at 10 m above ground
- $T = 100$ temporal samples used to calculate $\bar{\mathcal{S}}(\theta)$, resampled at every Gibbs-step to provide an unbiased estimation of the full 350640 slices in the dataset
- $K = 50$ ensemble members simulated via π_θ, industry standard in operational weather forecasting
- $M = 16$ proposals per variable, per Gibbs-step/SMC-step
- $N = 120$ total Gibbs-steps/SMC-steps, more than enough to witness convergence to a stationary distribution
- $B = 20$ burn-in steps discarded from the start of the posterior data (except for the Gibbs optimiser using $B = 35$)

Table 1: Experimental Design Hyperparameters

Parameter	**Value**
Atmospheric Variables (V)	5 (z500, t850, t2m, u10, v10)
Temporal Samples (T)	100 per step (resampled)
Ensemble Size (K)	50 members
Proposals per Variable (M)	16 per Gibbs/SMC step
Total Steps (N)	120 Gibbs/SMC steps
Burn-in Period (B)	20 steps (35 for Gibbs optimiser)
Dataset Size	350,640 temporal slices
Prior Distribution	Gamma(shape=2.0, scale=0.13)

4.2 Score Trajectories, Parameter Evolution and Posterior Distributions

All algorithmic variants appear to converge to a relatively stable score during the training phase. Trajectories show rapid improvement across the Gibbs and optimisation based algorithms, with SMC converging slower and to a more variable distribution. The convergence characteristics are summarized in Table 2.

Table 2: Algorithm Convergence Summary

Method	Score	Energy	CRPS	Conv. Steps	Quality
Gibbs	Energy	15.50	0.104	∼10	Stable
Gibbs	CRPS	15.50	0.101	∼5	Excellent
SMC	Energy	15.69	0.113	∼20	Good
SMC	CRPS	17.25	0.124	∼35	Moderate
Optimiser	Energy	15.43	0.105	∼15	Degenerate
Optimiser	CRPS	15.51	0.101	∼8	Good

Figure 5 shows the Gibbs-ABC posterior under the energy score discrepancy. The marginal parameter posteriors are broad. The score trajectory reaches a stable level by only the second iteration.

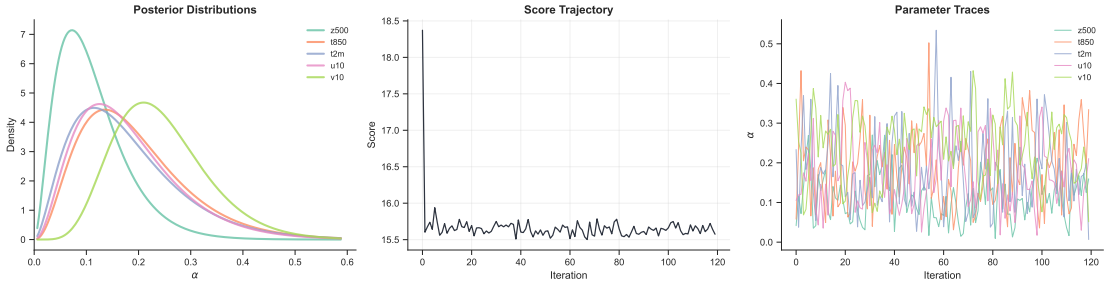

Figure 5: Gibbs-ABC with Energy Score: Posterior fitted parameter distributions [left], Energy score trajectory [centre], Parameter trace plots [right]

Figure 6 shows the Gibbs-ABC posterior under the CRPS discrepancy. Marginal posteriors are peaked, with less overlap between variables. Score trajectories stabilise rapidly within the first iterations, and parameter traces remain concentrated around consistent narrow ranges.

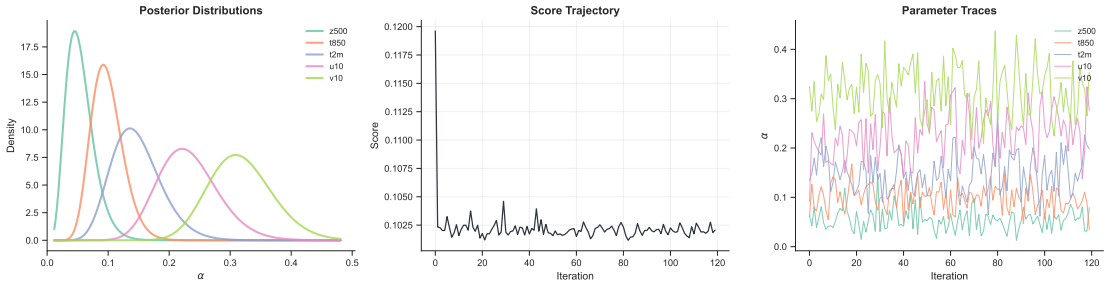

Figure 6: Gibbs-ABC with CRPS: Posterior fitted parameter distributions [left], CRPS score trajectory [centre], Parameter trace plots [right]

Figure 7 shows the SMC-ABC posterior under the energy score discrepancy. The energy score improves across the first few dozen SMC steps, with the adaptive proposal mechanism eventually finding the higher-mass areas of the posterior.

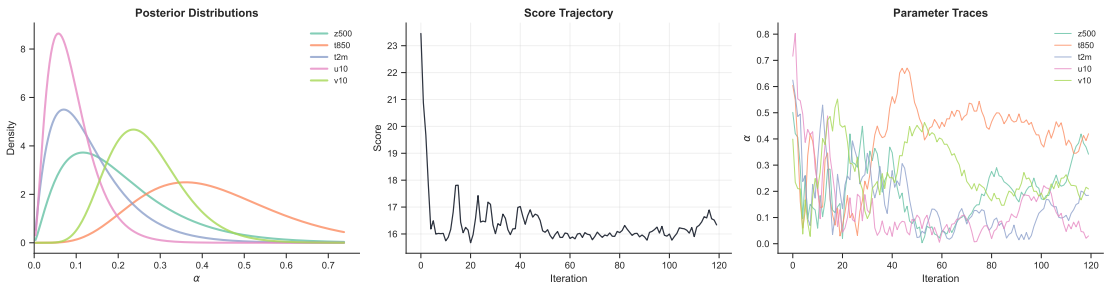

Figure 7: SMC-ABC with Energy Score: Posterior fitted parameter distributions [left], Energy score trajectory [centre], Parameter trace plots [right]

Figure 8 shows the SMC-ABC posterior under the CRPS discrepancy. The marginal parameter posteriors are wide; the CRPS appears unstable and wanders within a range.

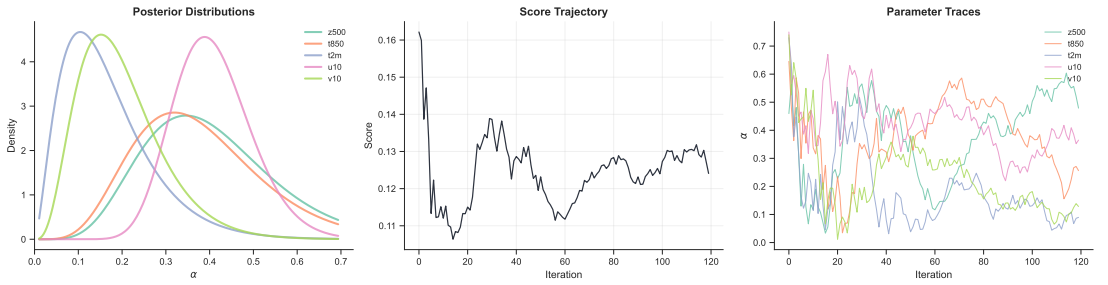

Figure 8: SMC-ABC with CRPS: Posterior fitted parameter distributions [left], CRPS trajectory [centre], Parameter trace plots [right]

Figure 9 shows the Gibbs optimisation posterior under the energy score. Marginals are

degenerate and score values plateau quickly.

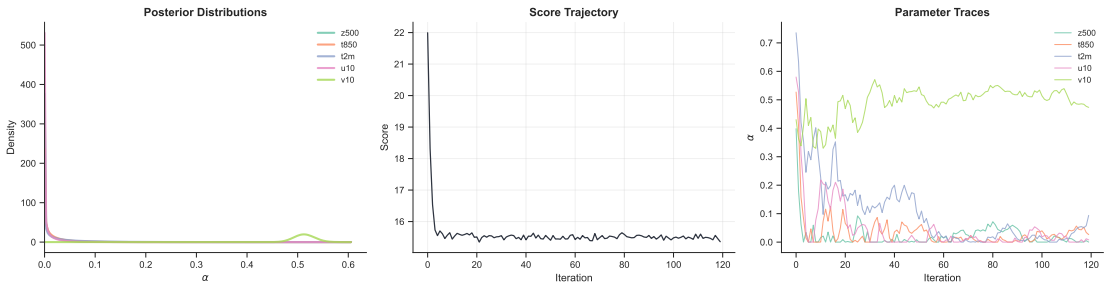

Figure 9: Gibbs optimisation with Energy Score: Posterior fitted parameter distributions [left], Energy score trajectory [centre], Parameter trace plots [right]

Figure 10 shows the Gibbs optimisation posterior under the CRPS. Parameter posterior marginal distributions are narrow and ordered; score values drop sharply in early iterations and stabilise. Trace plots indicate rapid convergence followed by small-scale variation.

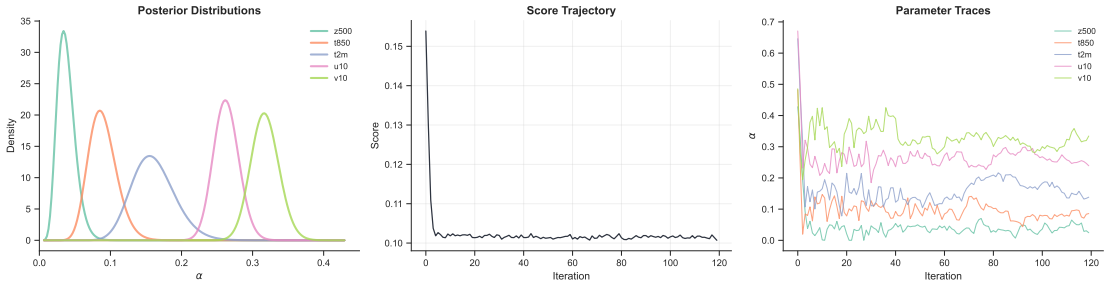

Figure 10: Gibbs optimisation with CRPS: Posterior fitted parameter distributions [left], CRPS trajectory [centre], Parameter trace plots [right]

4.3 Posterior Analysis

4.3.1 CRPS vs Energy Score

Across the algorithmic variants, the CRPS objective produces posteriors with sharper, more localised modes and clearer separation between variables. The energy score discrepancy yields generally broader marginals. This behaviour is consistent with the scoring rules' structural properties: the CRPS, applied marginally and subsequently aggregated, enforces per-variable calibration and penalises parameter values that degrade

performance for any individual variable, constraining each component more tightly. The multivariate energy score aggregates across variables and more easily allows for trade-offs between them, creating compensation effects where a deterioration in one component can be offset by improvement in another. Under the energy score, increases in perturbation amplitude for one variable can be balanced by decreases in another without materially altering the aggregate score, leading to a posterior landscape characterised by ridges and extended regions of near-constant score. While the Gibbs-ABC procedure results in parameters which exhibit a level of non-identifiability, it still achieves calibration in terms of the chosen score. This implies that the forecast variability induced by the different scale parameters is somewhat interchangeable.

The ordering of posterior means obtained under CRPS (see Table 4), $\alpha_{\mathrm{v10}} > \alpha_{\mathrm{u10}} > \alpha_{\mathrm{t2m}} > \alpha_{\mathrm{t850}} > \alpha_{\mathrm{z500}}$, is physically interpretable: variables with larger posterior perturbation scales are less predictable from a meteorological point of view. In particular, the surface wind components (u10 and v10) are sensitive to sub-gridscale geography and their behaviour is dictated by fast-growing instabilities in turbulent fluid dynamics. This produces strong local variability and smaller scale spatial and temporal correlations not adequately captured in the coarse deterministic model, meaning their associated uncertainty is greater. In contrast, the geopotential height (z500) and the temperature at 850 hPa (t850) are generally smoother and retain greater spatio-temporal coherence at the scales captured in the model, reducing the perturbation amplitude required for calibration. Score-ABC driven inference under the CRPS produces a posterior structure that is consistent with domain knowledge and physical expectations, reinforcing the practical interpretability of this framework as a calibration tool in high-dimensional forecasting problems.

4.3.2 Gibbs vs SMC vs Optimisation Algorithms

For both scoring rules, Gibbs-like updates tend to converge rapidly, producing stable posteriors within relatively few sweeps. This is due to the conditional updates focusing proposals into high-probability regions for each coordinate. SMC-ABC converges more slowly, especially under the CRPS, reflecting the gradual adaptation of the proposal mechanism and the resampling noise inherent to a particle-based system. Under CRPS, both the Gibbs and Optimisation schemes produce similarly ordered and identifiable posteriors, but the SMC algorithm produces a less coherent posterior and perhaps fails to converge completely. Under the energy score, the SMC implementation results in a

broader coverage of parameter space whereas the Gibbs algorithm exploits a narrower band of parameters. The optimisation algorithm appears to get "stuck" in a local optima and produces a degenerate distribution. This indicates further tuning of hyperparameters is necessary (such as the optimisation proposal mechanism).

4.3.3 Parameter Estimates and Forecast Scores

Despite differences in the convergence behaviour across implementations, each individual Score-ABC procedure did successfully calibrate the α RFP scale parameter for the atmospheric variables and achieved low, stable forecast scores. Two versions of 'uncalibrated' RFP forecasting methods have been included as a baseline forecast to demonstrate the benefit of calibration.Reference methods (climatology, persistence forecasting, deterministic forecast + heteroscedastic noise) are also included in Table 3 for comparison. Please see Section 3.4 for details of these methods.

Table 3: Comparison of forecasting methods across algorithms and scores

Method	**Energy Score**	**CRPS**
Reference Methods		
Deterministic Gaussian	15.49	0.096
Persistence	30.53	0.156
Climatology	44.13	0.292
RFP Uncalibrated Baselines		
Raw RFP ($\alpha = 1.0$)	29.21	0.221
Prior RFP (α drawn from prior)	17.23	0.118
RFP ABC-Calibrated (Posterior Mean)		
Gibbs-Energy	**15.50**	**0.104**
Gibbs-CRPS	**15.50**	**0.101**
Optimiser-Energy	**15.43**	**0.105**
Optimiser-CRPS	**15.51**	**0.101**
SMC-Energy	**15.69**	**0.113**
SMC-CRPS	**17.25**	**0.124**

Table 4: ABC Parameter Means and 90% Credible Intervals

Method	z500	t850	t2m	u10	v10
Gibbs (Energy)	0.114 [0.023, 0.206]	0.189 [0.050, 0.347]	0.185 [0.047, 0.369]	0.178 [0.049, 0.338]	0.243 [0.096, 0.357]
Gibbs (CRPS)	0.055 [0.023, 0.095]	0.099 [0.059, 0.140]	0.149 [0.090, 0.207]	0.226 [0.149, 0.311]	0.316 [0.242, 0.392]
SMC (Energy)	0.204 [0.037, 0.385]	0.399 [0.098, 0.597]	0.162 [0.023, 0.401]	0.139 [0.030, 0.439]	0.272 [0.146, 0.452]
SMC (CRPS)	0.372 [0.118, 0.564]	0.375 [0.147, 0.552]	0.182 [0.048, 0.449]	0.422 [0.270, 0.599]	0.229 [0.073, 0.467]
Optimiser (Energy)	0.022 [0.000, 0.063]	0.034 [0.000, 0.088]	0.103 [0.000, 0.353]	0.044 [0.000, 0.200]	0.490 [0.365, 0.545]
Optimiser (CRPS)	0.043 [0.010, 0.065]	0.098 [0.062, 0.140]	0.165 [0.111, 0.211]	0.266 [0.226, 0.298]	0.328 [0.282, 0.398]

Notes: Parameter means shown on first row, 90% credible intervals [5th, 95th percentiles] on second row for each method. All parameters are RFP scaling factors for meteorological variables.

4.4 Forecast Samples and Spatial Coherence

Unlike grid-point independent methods such as the deterministic + heteroscedastic noise model mentioned in Section 3.4.3, the RFP approach with ABC calibration produces a competitive probabilistic forecast whilst retaining intra-variable spatial correlation structure, essential for effective decision-making in real world forecasting applications.

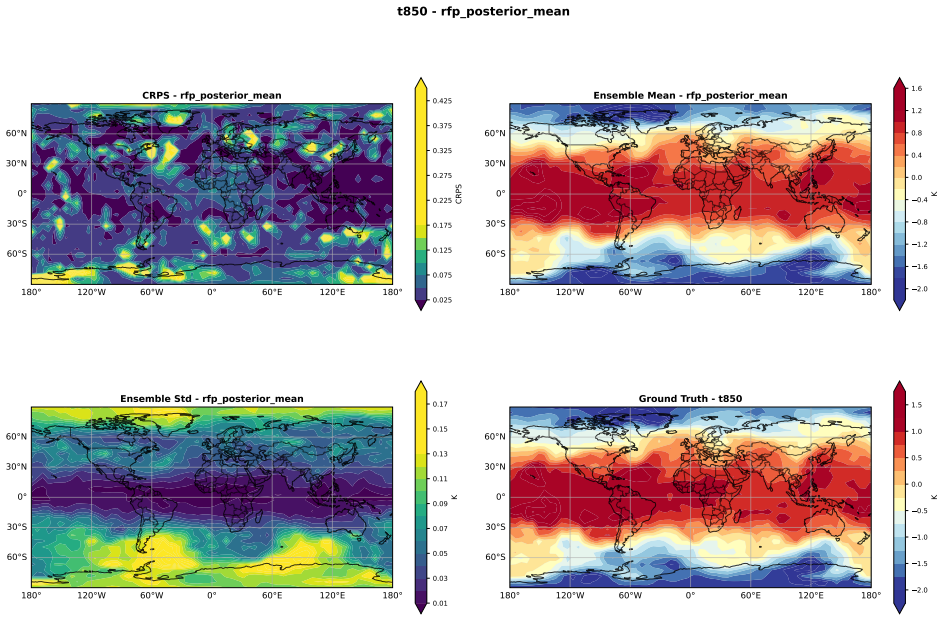

Figure 11: Spatial CRPS, the ground truth, ensemble mean and standard deviation resulting from a Gibbs-ABC (Energy) calibrated RFP ensemble for the variable t850

4.5 Computational Efficiency and Scalability

The batched implementation of the Gibbs-based Score-ABC algorithms, with efficiency mechanisms such as temporal resampling and CRPS approximation, successfully addressed the computational challenges associated with high-dimensional ABC schemes. Each procedure successfully completed the 120 Gibbs-steps within 12 hours on a single L40S GPU. It appears that score convergence was reached relatively early on in this process, meaning that the computational scale of the procedure can be feasibly reduced without excessive forecast performance degradation.

Memory management was a key aspect of the implementation requiring careful optimisation of batch sizes to maximise efficiency whilst avoiding memory errors during execution. A dynamic batch-size system was implemented to monitor system memory exhaustion and optimise forward-pass batch-sizes on-the-fly.

5 Discussion

The experimental results presented in Section 4 establish three primary findings:

1. Framework validity: Score-ABC converges and produces consistent results across different algorithmic implementations
2. RFP performance competitiveness: Score-ABC calibrated RFP methods are competitive with highly flexible and heteroscedastic methods such as the deterministic + heteroscedastic noise model, yet have the added benefit of spatial coherence
3. Practical viability: Score-ABC operates with efficiency despite implementation challenges (managing the dimensionality of both the forecast $\mathcal{X}$ and parameter Θ spaces)

The narrow performance variance across different implementations, combined with consistent improvements from both starting parameter draws and external simple baselines (climatology and persistence forecasts) show that using proper scoring rules as an ABC objective results in a mathematically-principled and practically effective procedure for forecast calibration and model parameter inference.

5.1 Limitations and Methodological Considerations

5.1.1 Computational Complexity and Scalability

The Score-ABC framework, like any ABC procedure, requires many forward model simulations. In particular, using Score-ABC to calibrate an ensemble forecasting model in the way described in this report requires substantial computational resources that may limit its applicability in some operational settings. Each parameter proposal necessitates generating an ensemble of K forecasts across T temporal samples, resulting in $K \times T$ forward passes through the deterministic model per proposal. With typical values of $K = 50$ ensemble members, $T = 100$ temporal samples, and $M = 16$ proposals per Gibbs step, a single variable update requires approximately 80,000 model evaluations. While GPU vectorisation and temporal resampling help to substantially reduce this burden, the computational cost is still high.

Even with the advantages over traditional NWP methods provided by MLWP, Score-ABC's computational burden may still be prohibitive for some models. The U-Net model used in this report takes milliseconds to conduct a forward-pass, but larger graph and transformer based models can take seconds to minutes (Lam et al., 2023) per evaluation. For these models, the computational time required for the implementation described in this report would be days or weeks, potentially exceeding available resources.

5.1.2 Temporal and Spatial Generalisation

The experimental design employs temporal resampling to reduce computational burden, evaluating forecast performance on randomly selected subsets of available data rather than the complete temporal sequence. While this provides an unbiased estimator of the population score, it may miss important temporal dependencies, seasonal effects or extreme events that could influence parameter calibration. The framework assumes that forecast quality is stationary across time, which may not hold for systems with strong cycles or climate trends. A more sophisticated temporal sampling system may lead to better generalisation performance.

Spatial aggregation also introduces potential limitations. The Score-ABC framework optimises global parameters across all spatial locations simultaneously, potentially averaging over important regional differences in forecast skill or uncertainty characteristics.

This global approach likely hides location-specific calibration needs, particularly in regions with distinct geographical characteristics. For example, forecasting in polar regions is generally associated with a higher level of uncertainty. This is clearly witnessed in the standard deviation of a Gibbs-CRPS calibrated ensemble forecast in Figure 11.

5.1.3 Scoring Rule Selection and Sufficiency

The choice of scoring rule fundamentally shapes the resulting posterior distribution, yet the framework provides limited guidance for selecting among competing scores. As demonstrated in the results, CRPS and energy score objectives yield markedly different parameter posteriors and identifiability characteristics. While both produce attractive forecasts, their differing approaches to marginal versus joint calibration creates an important methodological choice which requires a clear justification.

As mentioned in Section 3.1.1, proper scoring rules deliberately sacrifice the sufficiency property that traditional ABC summary statistics often aim to achieve. While this trade-off is strongly justified for the "deterministic model with stochastic augmentation" operational forecasting context, it means that Score-ABC posteriors reflect score-informed rather than likelihood-informed parameter distributions. This departure from conventional Bayesian principles may limit interpretability and comparability with alternative inference procedures.

5.1.4 Reference Table Approximation

The Gibbs-ABC implementation employs a reference table approach that selects the best-performing parameter from a small set of proposals rather than sampling from the true conditional posterior. While Clarté et al. (2021) demonstrate theoretical convergence properties for this approximation, it introduces a bias in finite samples toward posterior modes that may not truly reflect parameter uncertainty. This argmin selection mechanism particularly affects the optimisation-based Gibbs variant, which shows signs of getting trapped in local optima and produces degenerate distributions.

5.2 Future Research Directions and Implications

Whilst this report has demonstrated the viability of Score-ABC as a framework for calibrated probabilistic inference, there are extensive opportunities for further research in this topic. Algorithmic improvements are obvious next steps to build on the concepts discussed in this work. Including topics such as:

- Examination of other simulation methods within the Score-ABC framework, such as Hamiltonian or multiple chain Monte-Carlo algorithms
- Early rejection of clearly badly-performing parameter proposals before running a full scoring evaluation to reduce computational burden
- Increasing parameter sizes, in particular the number of proposals per variable M
- Exploring other scoring rules, including multi-objective rules to avoid trade-offs between score properties

Future work should also examine the theoretical properties of Score-ABC, such as the asymptotic and convergence behaviour of different algorithmic variants. As well as this, a more rigorous investigation of the behaviour of Score-ABC under model misspecification would significantly contribute to the theoretical justifications for using the framework.

Within the MLWP application, Score-ABC could be deployed to aid in the calibration of other uncertainty quantification models, such as RFP with regional parameters $\alpha_{\text{polar}}, \alpha_{\text{equatorial}}$. Beyond this domain, Score-ABC's principles apply broadly and should be explored in other fields such as epidemiology, economics and finance.

The demonstrated success and feasibility of Score-ABC in enabling calibrated probabilistic inference also highlights a significant development in MLWP: the staggering gain in computational efficiency of MLWP versus traditional NWP forecasting has truly opened the door to the wide topic of simulation-based methods which were previously infeasible (to even explore in a research setting, let alone deploy operationally). Further work should be done to establish the full range and extent of opportunities provided by simulation-based methods for the purposes of weather and climate forecasting.

6 Conclusion

This report has introduced Score-ABC, an approximate Bayesian computation framework for conducting likelihood-free parameter inference on calibrated probabilistic forecasting models. It addresses key challenges in probabilistic forecasting, a domain where models are often complex and likelihoods intractable or very computationally expensive (particularly in the case of uncertainty quantification for machine learning weather prediction models).

The main contribution of this work is demonstrating that proper scoring rules can serve as theoretically justified yet practical discrepancy measures within ABC schemes, as an alternative to more traditional constructed summary statistics and distance metrics. By aligning inference with forecast performance criteria, Score-ABC results in parameter posteriors which reflect adherence to proper scoring rule objectives, enforcing honest forecasting.

Experimental evaluation via calibrating the parameters of random field perturbations applied to a deterministic U-Net machine learning weather prediction model, establishes several key findings. Score and parameter convergence are witnessed under several algorithms and with different score objectives, producing spatially coherent and competitive forecasts. The difference in parameter posteriors generated under CRPS and Energy scores highlights the fact that the choice of scoring rule is a critical step in probabilistic forecast calibration, fundamentally shaping the posterior landscape and requiring careful attention. The framework's computational success demonstrates the feasibility of simulation-based inference in the MLWP domain and suggests great opportunities in exploring methods that were previously infeasible.

While limitations are present, this work establishes a foundation for future research into likelihood-free inference and probabilistic forecasting. Next steps may include theoretical work examining convergence and asymptotic guarantees, as well as algorithmic improvements to increase operational feasibility of Score-ABC type schemes.

To conclude, Score-ABC provides a theoretically justified and practically feasible framework for inference on calibrated probabilistic forecasting models. It represents a contribution to both approximate Bayesian computation literature and the probabilistic forecasting application in machine learning weather prediction, offering new tools for uncertainty quantification in complex systems.

7 Endmatter

7.1 Imperial College Research Computing Service

I would like to acknowledge computational resources and support provided by the Imperial College Research Computing Service (http://doi.org/10.14469/hpc/2232)

7.2 Code and Data Availability

All code used in this project is written in python and available at: `https://github.com/aturner22/unet_abc`. Environment and dependencies can be built using the UV package manager.

Coarse ERA5 data (Hersbach et al., 2020) can be downloaded via the work of Andrae et al. (2025) at `https://dataserv.ub.tum.de/index.php/s/m1524895` and converted to the correct format using the `create_dataset.py` script.

7.3 Large Language Model Usage Statement

Both Anthropic's Claude Sonnet and OpenAI's ChatGPT-4o/5 were used to aid with bug fixing and to create plotting utilities throughout this work's code development process. GPT-4o/5 was used to help craft this report's structure and to give direction on the rewording of verbose paragraphs during this report's editing phase. Output from LLMs was not copied verbatim for the writing of this report.

References

Andrae, M., Landelius, T., Oskarsson, J., and Lindsten, F. (2025). Continuous Ensemble Weather Forecasting with Diffusion models.

Bouallègue, Z. B., Clare, M. C. A., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Dramsch, J. S., Lang, S. T. K., Raoult, B., Rabier, F., Chevallier, M., Sandu, I., Dueben, P., Chantry, M., and Pappenberger, F. (2024). The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context. *Bulletin of the American Meteorological Society*, 105(6):E864–E883.

Brier, G. W. (1950). Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1):1–3.

Bülte, C. (2025). Discussion: Uncertainty Quantification for Data-Driven Weather Models.

Bülte, C., Horat, N., Quinting, J., and Lerch, S. (2025). Uncertainty quantification for data-driven weather models. *Artificial Intelligence for the Earth Systems*.

Clarté, G., Robert, C. P., Ryder, R. J., and Stoehr, J. (2021). Componentwise approximate Bayesian computation via Gibbs-like steps. *Biometrika*, 108(3):591–607.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474.

Gneiting, T. and Katzfuss, M. (2014). Probabilistic Forecasting. *Annual Review of Statistics and Its Application*, 1(1):125–151.

Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Hersbach, H. (2000). Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, 15(5):559–570.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049.

Karras, T., Aittala, M., Aila, T., and Laine, S. (2022). Elucidating the Design Space of

Diffusion-Based Generative Models.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*.

Lorenz, E., Hurka, J., Heinemann, D., and Beyer, H. G. (2009). Irradiance Forecasting for the Power Prediction of Grid-Connected Photovoltaic Systems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2(1):2–10.

Magnusson, L., Nycander, J., and Källén, E. (2009). Flow-dependent versus flow-independent initial perturbations for ensemble prediction. *Tellus A*, 61(2):194–209.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180.

Mittermaier, M. P. (2008). The Potential Impact of Using Persistence as a Reference Forecast on Perceived Forecast Skill. *Weather and Forecasting*, 23(5):1022–1031.

Murphy, A. H. (1973). A New Vector Partition of the Probability Score. *Journal of Applied Meteorology (1962-1982)*, 12(4):595–600.

Palmer, T. N. (2002). The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society*, 128(581):747–774.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., Ewalds, T., Stott, J., Mohamed, S., Battaglia, P., Lam, R., and Willson, M. (2025). Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer, New York, NY.

Siegert, S. (2017). Simplifying and generalising Murphy's Brier score decomposition. *Quarterly Journal of the Royal Meteorological Society*, 143(703):1178–1183.

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765.

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272.