

RESEARCH ARTICLE

Comparison of uncertainty quantification methods for cloud simulation

T. Janjić¹  | M. Lukáčová-Medvidová²  | Y. Ruckstuhl¹ | B. Wiebe²

¹Meteorological Institute,
Ludwig-Maximilians-University, Munich,
Germany

²Institute of Mathematics, Johannes
Gutenberg-University, Mainz, Germany

Correspondence

T. Janjić, Meteorological Institute,
Ludwig-Maximilians-University, Munich,
Germany.
Email: tijana.janjic@ku.de

Present address

Mathematical Institute for Machine
Learning and Data Science, Katholische
Universitaet, Eichstaett–Ingolstadt,
Eichstaett, Germany

Funding information

Deutsche Forschungsgemeinschaft,
Grant/Award Numbers: 257899354-TRR
165, Project A2, B6, JA1077/4-1; Mainz
Institut of Multiscale Modelling,
Grant/Award Number: M3odel;
Gutenberg Research College,
Grant/Award Number: JGUMainz

Abstract

Quantification of evolving uncertainties is required for both probabilistic forecasting and data assimilation in weather prediction. In current practice, the ensemble of model simulations is often used as a primary tool to describe the required uncertainties. In this work, we explore an alternative approach, the so-called stochastic Galerkin (SG) method, which integrates uncertainties forward in time using a spectral approximation in stochastic space. In an idealized two-dimensional model that couples nonhydrostatic weakly compressible Navier–Stokes equations to cloud variables, we first investigate the propagation of initial uncertainty. Propagation of initial perturbations is followed through time for all model variables during two types of forecast: the ensemble forecast and the SG forecast. Series of experiments indicate that differences in performance of the two methods depend on the system state and truncations used. For example, in more stable conditions, the SG method outperforms the ensemble of simulations for every truncation and every variable. However, in unstable conditions, the ensemble of simulations would need more than 100 members (depending on the model variable) and the SG method more than a truncation at five to produce comparable but not identical results. As estimates of the uncertainty are crucial for data assimilation, secondly we instigate the use of these two methods with the stochastic ensemble Kalman filter. The use of the SG method avoids evolution of a large ensemble, which is usually the most expensive component of the data assimilation system, and provides results comparable with the ensemble Kalman filter in the cases investigated.

KEYWORDS

convective scale, data assimilation, ensembles, forecasting (methods), stochastic Galerkin, uncertainty quantification

1 | INTRODUCTION

Quantifying the uncertainty of weather and climate predictions is necessary and of crucial importance, especially

for highly variable fields such as precipitation. Due to the chaotic atmosphere, probabilistic approaches to prediction have become standard in weather and climate forecasting, describing the evolution of initial condition uncertainty

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Quarterly Journal of the Royal Meteorological Society* published by John Wiley & Sons Ltd on behalf of the Royal Meteorological Society.

and its combination with model uncertainty, ultimately providing the end forecast uncertainty on various time and spatial scales.

The current method used for probabilistic forecasting utilizes an ensemble of initial conditions, which are then disturbed further by either perturbing relevant parameters, using stochastic parameterizations, or using dynamic perturbations (see (Palmer, (2019) for a recent overview). The predictability of a chaotic system is flow-dependent and contingent on the atmosphere's initial state. Some weather regimes may be highly volatile, while others may contain substantial predictability. Ensemble simulations provide a range of plausible forecasts, offering an estimate of trust in predictions and probabilities of possible outcomes. Richardson (2000) argues that additional information in ensemble prediction systems provides benefits to users equivalent to improvements in forecast models and assimilation systems over many years.

However, ensuring that the ensemble spread encompasses a true atmospheric state has been quite difficult. Uncertainties in the model itself and its biases lead to restricted sampling of the forecast space. Furthermore, numerical weather prediction models are reaching increasingly higher resolutions, even globally resolving physical processes that were previously unrepresented. To encompass more physical processes, the atmospheric state that we describe with these models will have more variables and parameters in the future. Therefore, the higher resolution will lead to increased nonlinearity of the system, a higher number of processes that are (partially) resolved, and a greater number of variables in the state vector that are highly non-Gaussian in nature (Gustafsson *et al.*, 2018; Janjić *et al.*, 2021). For example, for properly representing microphysical processes, variables such as rain, graupel, and snow, which are highly non-Gaussian, are required. Both increased nonlinearity, allowing for more complex dynamics, and an increase in the number of such prognostic variables would require more ensemble members to describe the non-Gaussian error statistics well. These effects require more ensemble members for the representation of uncertainty, (Ruckstuhl and Janjić, (2018)) and hence are even more costly in computational time.

We explore here an alternative approach, the so-called stochastic Galerkin (SG) method, for evolution of uncertainties forward in time. Uncertainty propagation in complex models, with the goal of systematically quantifying uncertainties and errors in models, simulations, and experiments and their impact on predicted quantities of interest, forms an active and innovative area of mathematical research. In recent years, a wide variety of uncertainty quantification methods have been proposed and investigated in the context of physical and engineering applications. These methods include SG

methods based on generalized polynomial chaos (Xiu and Karniadakis, 2002a; Wan and Karniadakis, 2006; Poëtte *et al.*, 2009; Tryoen *et al.*, 2010; Després *et al.*, 2013), stochastic collocation methods (Xiu and Hesthaven, 2005; Ma and Zabaras, 2009; Witteveen *et al.*, 2009), and multilevel Monte Carlo (MC) methods (Mishra and Schwab, 2012; Mishra *et al.*, 2013). Each of these method groups has its own pros and cons. While the results obtained by the ensemble (MC) simulations are generally very robust, the approach is not efficient due to the large number of simulations required (Elman *et al.*, 2011; Mishra *et al.*, 2013; Šukys *et al.*, 2013). As argued above, this approach will become extremely costly for future atmospheric applications. In the present article, we propose an alternative approach based on the spectral methodology in stochastic space that is efficient and accurate if the number of uncertain data/parameters remains moderate.

In this article, we consider and compare the standard MC method and SG to propagate uncertainties through a two-dimensional cloud model. In Section 2, we describe the model that we will use to illustrate similarities and differences between the two approaches and discuss how to consider the model's uncertainties. For this first comparison study, we chose a simple model in order to be able to compute many ensemble members and therefore calculate more accurate statistics. Section 3 describes two uncertainty quantification methods explored in our study: the MC (ensemble) and SG methods. The SG method integrates uncertainties forward in time using a spectral approximation in stochastic space (i.e., random direction). The methods are compared for different truncation numbers in stochastic space for SG and different numbers of ensemble members for MC in Section 4. In Section 5, the data assimilation results using a stochastic ensemble Kalman filter (EnKF) with and without SG, as well as with and without model error caused by unknown parameters, are compared. Conclusions from the study are drawn in Section 6.

2 | MATHEMATICAL MODEL AND ITS UNCERTAINTY

For our study, we use a two-dimensional mathematical model of cloud dynamics, which is based on the compressible nonhydrostatic Navier–Stokes equations for a moist atmosphere:

$$\begin{aligned} \rho'_t + \nabla \cdot (\rho \mathbf{u}) &= 0, \\ (\rho \mathbf{u})_t + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u} + p' \text{Id} - \mu_m \rho (\nabla \mathbf{u} + (\nabla \mathbf{u})^\top)) & \\ &= -\rho' g \mathbf{e}_2, \\ (\rho \theta)'_t + \nabla \cdot (\rho \theta \mathbf{u} - \mu_h \rho \nabla \theta) &= S_\theta, \end{aligned} \quad (\text{M1.a})$$

$$\begin{aligned}
 (\rho q_v)_t + \nabla \cdot (\rho q_v \mathbf{u} - \mu_q \rho \nabla q_v) &= \rho(-C + E), \\
 (\rho q_c)_t + \nabla \cdot (\rho q_c \mathbf{u} - \mu_q \rho \nabla q_c) &= \rho(C - A_1 - A_2), \\
 (\rho q_r)_t + \nabla \cdot (-v_q \rho q_r \mathbf{e}_2 + \rho q_r \mathbf{u} - \mu_q \rho \nabla q_r) \\
 &= \rho(A_1 + A_2 - E).
 \end{aligned}
 \tag{M1.b}$$

The model is a combination of Equation (M1.a), which realizes basic physical principles, and the equations for liquid water phases (Equation M1.b), which model warm cloud dynamics. We denote the time variable by t and the space vector by \mathbf{x} ; $\mathbf{x} = (x_1, x_2)$. For simplicity, we often omit the time and space dependence. We use the following notation: ρ is the density, $\mathbf{u} = (u, v)^\top$ is the velocity vector, θ is the moist potential temperature, p is the pressure, \mathbf{g} is the acceleration due to gravity, $\mathbf{e}_2 = (0, 1)^\top$. Further, μ_m is the dynamic viscosity, μ_h is the thermal conductivity, μ_q is the cloud diffusivity. In the simulations presented below we choose $\mu_m = 10^{-3}$ Pa·s and $\mu_h = \mu_q = 10^{-2}$ Pa·s, but other physically relevant values can be used as well. The dynamics of interest is described by a perturbation $\rho', \mathbf{u}, (\rho\theta)'$ of a background state (the hydrostatic equilibrium) to avoid numerical instabilities solving the Navier–Stokes equations (M1.a) in a weakly compressible regime. Thus, we use primes to denote perturbations of hydrostatic equilibrium. For a detailed description of the model, we refer the reader to Chertock *et al.* (2019). The cloud variables representing the mass concentration of water vapor, cloud droplets, and rain drops, $q_v, q_c,$ and $q_r,$ respectively, are given by

$$q_\ell = \frac{\text{mass of the respective phase}}{\text{mass of dry air}} \quad \text{for } \ell \in \{v, c, r\}.$$

The terms C and E represent phase changes between vapor and cloud water (droplets), A_1 and A_2 represent collision processes, which lead to the formation of large droplets and thus precipitation, and v_q is the raindrop fall velocity.

Note that the systems (M1.a) and (M1.b) are coupled through the source term S_θ , which represents the impact of phase changes and is given by

$$S_\theta = \rho \frac{L\theta}{c_p T} (C - E).$$

For a detailed description of S_θ and the terms $E, C, A_1, A_2,$ and $v_q,$ see Chertock *et al.* (2019). The temperature T can be obtained from the moist adiabatic ideal gas equation:

$$T = \frac{R}{R_m} \theta \left(\frac{p}{p_0} \right)^{R_m/c_p},$$

where $p_0 = 10^5$ Pa is the reference pressure at sea level. In addition to the usual definition of potential temperature, we use $R_m = (1 - q_v - q_c - q_r)R + q_v R_v,$ with the ideal gas constant of dry air $R = 287.05 \text{ J} \cdot \text{kg}^{-1} \cdot \text{K}^{-1},$ the

gas constant of water vapor $R_v = 461.51 \text{ J} \cdot \text{kg}^{-1} \cdot \text{K}^{-1},$ and the specific heat capacity of dry air at constant pressure $c_p = 1005 \text{ J} \cdot \text{kg}^{-1} \cdot \text{K}^{-1}.$ In order to close the system, we determine the pressure from the equation of state that includes moisture:

$$p = p_0 \left(\frac{R\rho\theta}{p_0} \right)^{\gamma_m} \quad \text{with} \quad \gamma_m = \frac{c_p}{c_p - R_m}. \tag{2}$$

We note that in the dry case, that is, when $q_v = q_c = q_r = 0,$ R_m reduces to $R, S_\theta = 0,$ and the moist ideal gas equation as well as the moist equation of state becomes its dry analog.

In Figure 1 we depict a selection of the modeled fields in the setup of the well-known meteorological benchmark describing the free convection of a smooth warm air bubble (see, e.g., (Davies and Taylor, 1950; Bryan and Fritsch, 2002)).

Meteorological applications typically inherit several sources of uncertainties, such as model parameters, initial, and boundary conditions. Consequently, purely deterministic models are insufficient in such situations and more sophisticated methods need to be applied to analyze the influence of uncertainties on numerical solutions. In general, there are different ways to represent and take into account a model’s uncertainty. In this article, we choose a widely used approach, namely to describe the uncertainty by a random variable $\omega.$ We will consider the case of a single random input and leave the case with multiple random input parameters for future research.

We apply a widely used approach to represent the uncertainty by means of random fields. To this end we define an abstract probability space (Γ, Σ, P) and denote an outcome by $\omega, \omega \in \Gamma.$ We assume that the initial data are random variables depending on $\omega,$ that is,

$$(\rho q_v) \Big|_{t=0} = (\rho q_v)(\mathbf{x}, t = 0, \omega). \tag{3}$$

Consequently, the solution at later time is also a vector of random variables depending on $\omega,$ that is, $(\rho q_\ell)(\mathbf{x}, t, \omega)$ for $\ell \in \{v, c, r\}.$ Due to the coupling of the source term S_θ in the energy equation, the fluid components are also random variables that depend on $\omega,$ that is,

$$\rho'(\mathbf{x}, t, \omega), (\rho \mathbf{u})(\mathbf{x}, t, \omega) \quad \text{and} \quad (\rho\theta)'(\mathbf{x}, t, \omega).$$

Thus, the fully stochastic cloud model is a system of random partial differential equations given through

$$\begin{aligned}
 (\rho'(\omega))_t + \nabla \cdot ((\rho \mathbf{u})(\omega)) &= 0, \\
 ((\rho \mathbf{u})(\omega))_t + \nabla \cdot ((\rho \mathbf{u})(\omega) \otimes \mathbf{u}(\omega) + p'(\omega) \text{Id} \\
 - \mu_m \rho(\omega) (\nabla \mathbf{u}(\omega) + (\nabla \mathbf{u}(\omega))^\top)) &= -\rho'(\omega) \mathbf{g} \mathbf{e}_2, \\
 ((\rho\theta)'(\omega))_t + \nabla \cdot ((\rho\theta)(\omega) \mathbf{u}(\omega) - \mu_h \rho(\omega) \nabla \theta(\omega)) &= S_\theta(\omega),
 \end{aligned}
 \tag{M2.a}$$

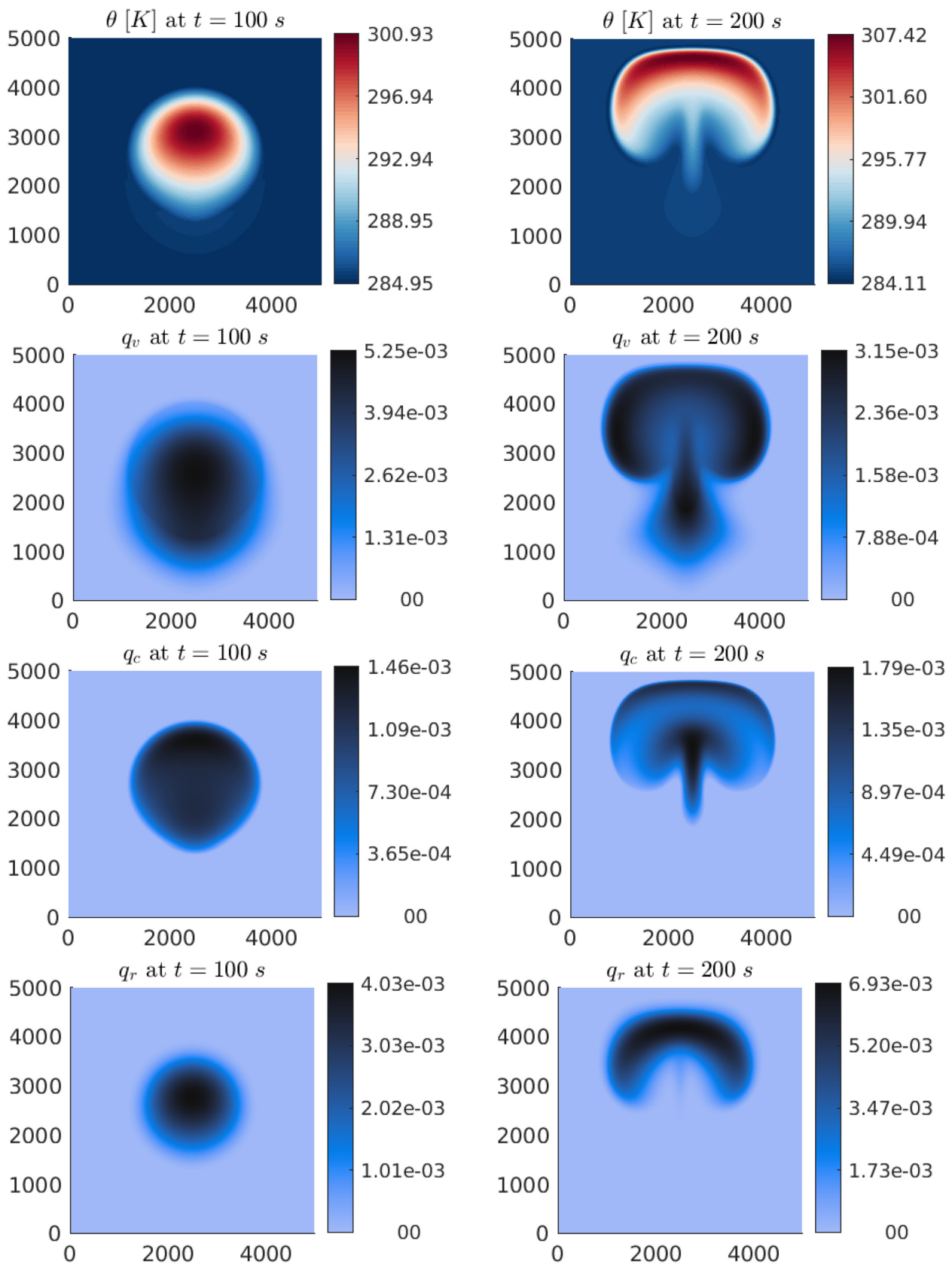


FIGURE 1 Potential temperature θ , water vapor concentration q_v , cloud drop concentration q_c , and rain concentration q_r at times $t = 100$ (left column) and 200 s (right column) simulated on a 320×320 mesh. [Colour figure can be viewed at [wileyonlinelibrary.com](https://onlinelibrary.wiley.com/terms-and-conditions)]

$$\begin{aligned}
& ((\rho q_v)(\omega))_t + \nabla \cdot ((\rho q_v)(\omega)\mathbf{u}(\omega) - \mu_q \rho(\omega) \nabla q_v(\omega)) \\
& = \rho(\omega)(-C + E)(\omega), \\
& ((\rho q_c)(\omega))_t + \nabla \cdot ((\rho q_c)(\omega)\mathbf{u}(\omega) - \mu_q \rho(\omega) \nabla q_c(\omega)) \\
& = \rho(\omega)(C - A_1 - A_2)(\omega), \\
& ((\rho q_r)(\omega))_t + \nabla \cdot ((\rho q_r)(\omega)(-v_q(\omega)\mathbf{e}_2 + \mathbf{u}(\omega)) \\
& - \mu_q \rho(\omega) \nabla q_r(\omega)) = \rho(\omega)(A_1 + A_2 - E)(\omega). \quad (\text{M2.b})
\end{aligned}$$

Note that it is also possible to start with additional uncertain initial fluid variables or even uncertain parameters which will lead to the same fully stochastic system (M2). The system (M2) can be understood in the so-called distributional sense. It means that we do not require that the equations hold for each (t, \mathbf{x}, ω) . It suffices that an integral (weak) formulation holds for almost all $t \in [0, T]$ and almost surely in ω . Consequently, we consider the so-called weak (distributional) solution in physical space/time. With respect to $\omega \in \Gamma$, we speak about a (stochastically) strong solution. It can be proved that the numerical methods presented below converge to such distributional solutions, at least for simplified model problems. We refer the reader to our recent works, where we have proved the convergence of finite-volume uncertainty quantification methods for the compressible Navier–Stokes equations (Feireisl *et al.*, 2022; Feireisl and Lukáčová-Medviděová, 2023) and convergence and error estimates of the mixed finite-element–finite-volume method for the Navier–Stokes equations with potential temperature transport (M1.a) (Lukáčová-Medviděová and Schömer, 2022; Lukáčová-Medviděová and Schömer, 2023).

3 | UNCERTAINTY QUANTIFICATION METHODS

In this section, we briefly introduce two uncertainty quantification approaches that we will be using in our study, the MC (ensemble) approach and the SG method. For a more detailed description of the SG method applied to model (M2), we refer the reader to (Chertock *et al.*, 2019; Wiebe, 2021; Chertock *et al.*, 2023).

3.1 | The MC method

The well-known MC method belongs to a group of sampling methods that are widely used to propagate uncertainties in nonlinear problems. This sampling technique is nonintrusive, in the sense that existing numerical model codes do not need to be modified, and is very intuitive, as randomly sampling from input distributions constructs an ensemble from which response statistics can be

computed. Additionally, the efficiency is independent of the number of parameters, since one can sample simultaneously from many parameter distributions. A drawback of sampling methods is that they typically exhibit relatively slow convergence rates, for the MC method of the order $1/\sqrt{N}$ toward the expected value and variance, where N is the number of simulations. This is due to the solely statistical nature of the method and the fact that it does not exploit regularity associated with the parameter space. Thus, a large number of response realizations are required to construct a reasonable statistical ensemble. The Latin hypercube (Stein, 1987; Loh, 1996), and quasi-MC sampling methods (Niederreiter, 1992; Morokoff and Caflisch, 1995) exhibit higher convergence rates, but they are often infeasible for computationally complex problems such as applications involving coupled physics.

3.2 | The SG method

The SG method belongs to the group of stochastic spectral methods, the objective of which is to reduce significantly the number of deterministic model simulations required to construct moments of quantities of interest. The SG method represents uncertain inputs in a way that facilitates the evaluation of moments and distributions of the quantities of interest. This is achieved by employing spectral expansions in stochastic space: one projects the governing equations onto a finite-dimensional subspace spanned by the polynomial basis functions and ends up with deterministic equations for the spectral expansion coefficients. The choice of polynomials depends on the probability distribution of the random variable ω . A favorable choice is the utilization of polynomials that are orthogonal with respect to the probability density function, which in the following we will denote by μ . This choice yields the best error constant. The latter stands in the error estimate in front of $1/M^p$, where M denotes the number of basis polynomials and p the convergence order related to the smoothness (regularity) of the exact solution. Indeed, if the latter is infinitely differentiable with respect to the random variable, then the estimate holds for any $p \geq 1$ and we obtain an exponential convergence rate. In this article, we will assume a normal distribution, which implies that the use of the following Hermite polynomials is an optimal choice (Xiu and Karniadakis, 2002b):

$$\begin{aligned}
\Phi_k(\omega) &= 2^{-k/2} H_k \left(\frac{\omega - \mu_H}{\sqrt{2}\sigma_H} \right) \quad \text{with} \\
H_k(\omega) &= (-1)^k e^{\omega^2} \frac{d^k}{d\omega^k} e^{-\omega^2}, \quad (4)
\end{aligned}$$

where μ_H and σ_H are the mean value and standard deviation of probability distribution μ . This implies that the following orthogonal property holds:

$$\langle \Phi_k(\omega), \Phi_{k'}(\omega) \rangle = \int_{\Gamma} \Phi_k(\omega) \Phi_{k'}(\omega) \mu(\omega) d\omega = k! \delta_{kk'} \quad \text{for } 0 \leq k, k' \leq M, \quad (5)$$

where $\delta_{kk'}$ is the Kronecker symbol, $\Gamma = \mathbb{R}$, and μ is the corresponding probability density function.

Each variable of the model will be approximated using M basis polynomials, as follows:

$$\begin{aligned} \rho'(\mathbf{x}, t, \omega) &\approx \sum_{k=0}^M (\hat{\rho}')_k(\mathbf{x}, t) \Phi_k(\omega), \quad \rho \mathbf{u}(\mathbf{x}, t, \omega) \\ &\approx \sum_{k=0}^M (\hat{\rho} \mathbf{u})_k(\mathbf{x}, t) \Phi_k(\omega), \quad (\rho \theta)'(\mathbf{x}, t, \omega) \\ &\approx \sum_{k=0}^M ((\hat{\rho} \theta)')_k(\mathbf{x}, t) \Phi_k(\omega) \end{aligned} \quad (6)$$

and

$$\rho q_{\ell}(\mathbf{x}, t, \omega) \approx \sum_{k=0}^M (\hat{\rho} q_{\ell})_k(\mathbf{x}, t) \Phi_k(\omega) \quad \text{with } \ell \in \{v, c, r\}. \quad (7)$$

To describe the SG method in more detail, we follow Chertock *et al.* (2023). Projecting (M2) onto the finite-dimensional space spanned by the Hermite basis polynomials, we obtain, for $k = 0, \dots, M$,

$$\begin{aligned} \langle (\rho')_t + \nabla \cdot (\rho \mathbf{u}), \Phi_k \rangle &= 0, \\ \langle (\rho \mathbf{u})_t + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u} + p' \text{Id} - \mu_m \rho (\nabla \mathbf{u} + (\nabla \mathbf{u})^T)), \Phi_k \rangle \\ &= \langle -\rho' g \mathbf{e}_d, \Phi_k \rangle, \\ \langle (\rho \theta)'_t + \nabla \cdot (\rho \theta \mathbf{u} - \mu_h \rho \nabla \theta), \Phi_k \rangle &= \langle S_{\theta}, \Phi_k \rangle, \end{aligned}$$

and

$$\begin{aligned} \langle (\rho q_v)_t + \nabla \cdot (\rho q_v \mathbf{u} - \mu_q \rho \nabla q_v), \Phi_k \rangle &= \langle \rho(-C + E), \Phi_k \rangle, \\ \langle (\rho q_c)_t + \nabla \cdot (\rho q_c \mathbf{u} - \mu_q \rho \nabla q_c), \Phi_k \rangle \\ &= \langle \rho(C - A_1 - A_2), \Phi_k \rangle, \\ \langle (\rho q_r)_t + \nabla \cdot (\rho q_r(-v_q \mathbf{e}_d + \mathbf{u}) - \mu_q \rho \nabla q_r), \Phi_k \rangle \\ &= \langle \rho(A_1 + A_2 - E), \Phi_k \rangle. \end{aligned}$$

Now, substituting (6) and (7) and using the orthogonality property (5) yields $M + 1$ deterministic equations for each previous equation:

$$\begin{aligned} ((\hat{\rho}')_k)_t + \nabla \cdot (\hat{\rho} \mathbf{u})_k &= 0, \\ ((\hat{\rho} \mathbf{u})_k)_t + \nabla \cdot (\hat{\mathbf{u}}_k + (\hat{p}')_k \text{Id}) - \mu_m (\hat{\mathbf{d}}_1)_k &= -(\hat{\rho}')_k g \mathbf{e}_d, \\ ((\hat{\rho} \theta)'_k)_t + \nabla \cdot (\hat{\theta} (\hat{\rho} \mathbf{u})_k + \hat{\eta}_k) - \mu_h (\hat{\mathbf{d}}_2)_k &= (\hat{S}_{\theta})_k, \end{aligned} \quad (8)$$

and

$$\begin{aligned} ((\hat{\rho} q_v)_k)_t + \nabla \cdot ((\hat{\eta}_1^q)_k) - \mu_q (\hat{\mathbf{d}}_1^q)_k &= (\hat{r}_1)_k, \\ ((\hat{\rho} q_c)_k)_t + \nabla \cdot ((\hat{\eta}_2^q)_k) - \mu_q (\hat{\mathbf{d}}_2^q)_k &= (\hat{r}_2)_k, \\ ((\hat{\rho} q_r)_k)_t + \nabla \cdot ((\hat{\eta}_3^q)_k) - \mu_q (\hat{\mathbf{d}}_3^q)_k &= (\hat{r}_3)_k, \end{aligned} \quad (9)$$

for $k = 0, \dots, M$. The coefficients that appear in Equations (8) and (9) are defined as follows:

$$\begin{aligned} \rho \mathbf{u} \otimes \mathbf{u} &= \sum_{k=0}^M \hat{\mathbf{u}}_k \Phi_k, \quad p' = \sum_{k=0}^M (\hat{p}')_k \Phi_k, \\ \nabla \cdot (\rho (\nabla \mathbf{u} + (\nabla \mathbf{u})^T)) &= \sum_{k=0}^M (\hat{\mathbf{d}}_1)_k \Phi_k, \\ \theta' \rho \mathbf{u} &= \sum_{k=0}^M \hat{\eta}_k \Phi_k, \quad \nabla \cdot (\rho \nabla \theta) = \sum_{k=0}^M (\hat{\mathbf{d}}_2)_k \Phi_k, \\ S_{\theta} &= \sum_{k=0}^M (\hat{S}_{\theta})_k \Phi_k, \\ \rho q_v \mathbf{u} &= \sum_{k=0}^M (\hat{\eta}_1^q)_k \Phi_k, \quad \nabla \cdot (\rho \nabla q_v) = \sum_{k=0}^M (\hat{\mathbf{d}}_1^q)_k \Phi_k, \\ \rho(-C + E) &= \sum_{k=0}^M (\hat{r}_1)_k \Phi_k, \\ \rho q_c \mathbf{u} &= \sum_{k=0}^M (\hat{\eta}_2^q)_k \Phi_k, \quad \nabla \cdot (\rho \nabla q_c) = \sum_{k=0}^M (\hat{\mathbf{d}}_2^q)_k \Phi_k, \\ \rho(C - A_1 - A_2) &= \sum_{k=0}^M (\hat{r}_2)_k \Phi_k, \\ \rho q_r \mathbf{u} - \rho q_r v_q \mathbf{e}_d &= \sum_{k=0}^M (\hat{\eta}_3^q)_k \Phi_k, \\ \nabla \cdot (\rho \nabla q_r) &= \sum_{k=0}^M (\hat{\mathbf{d}}_3^q)_k \Phi_k, \\ \rho(A_1 + A_2 - E) &= \sum_{k=0}^M (\hat{r}_3)_k \Phi_k, \end{aligned}$$

and are computed via discrete and inverse discrete Hermite transforms, which work in a manner analogous to the well-known Fourier transforms.

Note that Equations (8) and (9) mean that one needs to solve $M + 1$ equations for each previous equation in (M2) and that these coefficients in the equations are still dependent on both space and time. However, in contrast to the ensemble approach, M can be chosen much smaller than N , since spectral methods converge faster than sampling ones. Indeed, the spectral methods have an exponential convergence rate with respect to M , while the rate of the MC method is \sqrt{N} , where N denotes the size of the ensemble set. Furthermore, the choice of M does not depend on

the size of the atmospheric state vector, but rather on how well we can represent the distribution of dynamical variables with M components. The resulting system for the expansion coefficients then has to be solved by an accurate and efficient spatio-temporal approximation. Since the structure of the system of expansion coefficients is similar to that of the deterministic cloud model (Equation 1), one can adapt the deterministic numerical methods used for the space and time approximation in a suitable way. For more details on the resulting system and its approximation, we refer the reader to Chertock *et al.* (2019), Wiebe (2021), Chertock *et al.* (2023). From the coefficients obtained, one can compute the stochastic solution for each event ω by evaluating the respective expansions in Equations (6) and (7) and by using the Hermite polynomials from Equation (4). Also, a statistical and sensitivity analysis can be performed directly with minimal computational effort from the coefficients. For example, the mean and variance of $\rho q_\ell(\mathbf{x}, t, \omega)$, $\ell \in \{v, c, r\}$ are as follows.

• *Mean:*

$$\begin{aligned} \mathbb{E}[\rho q_\ell(\mathbf{x}, t, \omega)] &\approx \int_{\mathbb{R}} \sum_{k=0}^M (\widehat{\rho q_\ell})_k(\mathbf{x}, t) \Phi_k(\omega) \mu(\omega) \, d\omega \\ &= \sum_{k=0}^M (\widehat{\rho q_\ell})_k(\mathbf{x}, t) \int_{\mathbb{R}} \Phi_k(\omega) 1 \mu(\omega) \, d\omega \\ &= \sum_{k=0}^M (\widehat{\rho q_\ell})_k(\mathbf{x}, t) \int_{\mathbb{R}} \Phi_k(\omega) \Phi_0(\omega) \mu(\omega) \, d\omega \\ &= (\widehat{\rho q_\ell})_0(\mathbf{x}, t). \end{aligned} \tag{10}$$

• *Variance:*

$$\begin{aligned} \text{Var}(\rho q_\ell(\mathbf{x}, t, \omega)) &= \mathbb{E}[(\rho q_\ell(\mathbf{x}, t, \omega) - \mathbb{E}[\rho q_\ell(\mathbf{x}, t, \omega)])^2] \\ &\approx \mathbb{E}[(\rho q_\ell(\mathbf{x}, t, \omega) - (\widehat{\rho q_\ell})_0(\mathbf{x}, t))^2] \\ &\approx \mathbb{E}\left[\left(\sum_{k=1}^M (\widehat{\rho q_\ell})_k(\mathbf{x}, t) \Phi_k(\omega)\right)^2\right] \\ &= \int_{\mathbb{R}} \left(\sum_{k=1}^M (\widehat{\rho q_\ell})_k(\mathbf{x}, t) \Phi_k(\omega)\right) \\ &\quad \times \left(\sum_{j=1}^M (\widehat{\rho q_\ell})_j(\mathbf{x}, t) \Phi_j(\omega)\right) \mu(\omega) \, d\omega \\ &= \sum_{k=1}^M \sum_{j=1}^M (\widehat{\rho q_\ell})_k(\mathbf{x}, t) (\widehat{\rho q_\ell})_j(\mathbf{x}, t) \int_{\mathbb{R}} \Phi_k(\omega) \Phi_j(\omega) \mu(\omega) \, d\omega \\ &= \sum_{k=1}^M ((\widehat{\rho q_\ell})_k(\mathbf{x}, t))^2 k!. \end{aligned} \tag{11}$$

Higher-order moments and correlation functions can then be computed in a similar manner.

4 | FORECAST UNCERTAINTIES

In order to compare these two uncertainty quantification methods, we first consider the evolution of forecast uncertainties using Equation (M2). The initial conditions of the SG and MC simulations are chosen such that their respective probability density functions are identical. Since uncertainties in water and all of its phases are considered as one of the largest uncertainties in numerical weather prediction (NWP), we chose to start with a perturbation in q_v , caused, for example, by an uncertain parameter, and to consider how this uncertainty propagates to other variables in our model. We start our simulations with 20% normally distributed error in q_v , that is, for SG,

$$q_v(\mathbf{x}, t = 0, \omega) = (\widehat{q}_v)_0(\mathbf{x}, 0) + 0.2\omega(\widehat{q}_v)_0(\mathbf{x}, 0),$$

where $\omega \in \mathbb{R}$. Correspondingly, MC simulations start with an ensemble of N independent and identically distributed (i.i.d.) initial data:

$$q_v^i(\mathbf{x}, t = 0, \omega^i) = (\widehat{q}_v)_0(\mathbf{x}, 0) + 0.2\omega^i(\widehat{q}_v)_0(\mathbf{x}, 0),$$

$$i = 1, \dots, N,$$

ensuring the statistics of q_v is consistent between the two methods. The initial conditions of the other variables are assumed to be known and are therefore deterministic, yielding the following initial conditions.

• *SG:* For the cloud variables we use

$$\begin{aligned} (\widehat{q}_v)_0(\mathbf{x}, 0) &= 0.005 (\widehat{\theta}^v)_0(\mathbf{x}, 0), \\ (\widehat{q}_v)_1(\mathbf{x}, 0) &= 0.2 (\widehat{q}_v)_0(\mathbf{x}, 0), \\ (\widehat{q}_v)_k(\mathbf{x}, 0) &= 0 \quad \text{for } 2 \leq k \leq M, \\ (\widehat{q}_c)_0(\mathbf{x}, 0) &= 10^{-4} (\widehat{\theta}^c)_0(\mathbf{x}, 0), \quad (\widehat{q}_c)_k(\mathbf{x}, 0) = 0 \\ &\quad \text{for } 1 \leq k \leq M, \\ (\widehat{q}_r)_0(\mathbf{x}, 0) &= 10^{-6} (\widehat{\theta}^r)_0(\mathbf{x}, 0), \quad (\widehat{q}_r)_k(\mathbf{x}, 0) = 0 \\ &\quad \text{for } 1 \leq k \leq M, \end{aligned}$$

and for the Navier–Stokes variables

$$\begin{aligned} (\widehat{\rho}^v)_0(\mathbf{x}, 0) &= -\bar{\rho}(\mathbf{x}) \frac{(\widehat{\theta}^v)_0(\mathbf{x}, 0)}{\bar{\theta}(\mathbf{x}) + (\widehat{\theta}^v)_0(\mathbf{x}, 0)}, \\ (\widehat{\rho}^v)_k(\mathbf{x}, 0) &= 0 \quad \text{for } 1 \leq k \leq M, \\ (\widehat{\rho \mathbf{u}})_k(\mathbf{x}, 0) &= 0 \quad \text{for } 0 \leq k \leq M, \\ ((\widehat{\rho \theta}^v))_0(\mathbf{x}, 0) &= \bar{\rho}(\mathbf{x}) (\widehat{\theta}^v)_0(\mathbf{x}, 0) + \bar{\theta} (\widehat{\rho}^v)_0(\mathbf{x}, 0) \\ &\quad + (\widehat{\theta}^v)_0(\mathbf{x}, 0) (\widehat{\rho}^v)_0(\mathbf{x}, 0), \\ ((\widehat{\rho \theta}^v))_k(\mathbf{x}, 0) &= 0 \quad \text{for } 1 \leq k \leq M, \end{aligned}$$

where

$$(\hat{\theta}')_0(\mathbf{x}, 0) = \begin{cases} 2\cos^2\left(\frac{\pi r}{2}\right), & r := \sqrt{(x_1 - 2500)^2 + (x_2 - 2000)^2} \\ & \leq 2000, \\ 0, & \text{otherwise,} \end{cases}$$

$$(\hat{\theta}')_k(\mathbf{x}, 0) = 0 \quad \text{for } 1 \leq k \leq M.$$

The background state is given through $\bar{\theta} = 285$ K, $p_0 = \bar{p} = 10^5$ Pa, and

$$\bar{\rho}(\mathbf{x}) = \frac{p_0}{R\bar{\theta}(\mathbf{x})} \pi_e(\mathbf{x})^{1/(\gamma-1)}, \quad \pi_e(\mathbf{x}) = 1 - \frac{gx_2}{c_p\bar{\theta}},$$

with $c_p = 1005$ J · kg⁻¹ · K⁻¹, $c_v = 718$ J · kg⁻¹ · K⁻¹, and $\gamma = c_p/c_v$.

- MC: We compute the initial conditions by evaluating the expansions in Equations (6), (7) for the above SG initial coefficients and samples ω^i , $i = 1, \dots, N$. Note that, for all variables except q_v , the initial conditions do not depend on ω and will therefore have the same value across the ensemble.

Note that we chose a normal distribution to represent an uncertain parameter in the initial conditions, since this is the usual choice for probabilistic forecasting and data assimilation. To avoid negative concentrations, we dismiss all MC samples starting with a negative water vapor concentration, and in SG we also dismiss all evaluations at quadrature points having a negative concentration. In general, we have not observed that this leads to major errors in our experiments, since the mean and standard deviation of the normal distribution were chosen in such a way that negative concentrations only occur with low probability.

We start in this experiment with nonzero values for the cloud drop concentration q_c and the rain concentration q_r to avoid values close to machine precision, since we want to investigate the errors. Furthermore, we apply the no-slip boundary conditions for the velocities and zero Neumann boundary conditions for the remaining variables, that is, $\nabla \rho' \cdot \mathbf{n} = 0$, $\nabla(\rho\theta)' \cdot \mathbf{n} = 0$, $\nabla(\rho q_\ell)' \cdot \mathbf{n} = 0$, $\ell \in \{v, c, r\}$.

The simulations are performed on 200×200 grid points in a squared domain $[0, 5000] \times [0, 5000]$ m², and for time $[0, 200]$ s, without data assimilation for now, which is discussed later in Section 5.

In Figure 2, the mean and standard deviations of the mass densities of water vapor ρq_v , cloud drops ρq_c , and rain ρq_r at time $t = 200$ s at each grid point of the domain are shown, simulated with the MC method using 50 and 10⁵ members and the SG method using two and 10 basis polynomials. Figure 3 illustrates the evolution of the averaged

sum of standard deviations (trace scaled with number of grid points) with time for MC simulations of 50 and 10⁵ members and SG with one, five, and 10 basis polynomials. As exhibited in Figures 2 and 3, uncertainty as represented with mean, standard deviation, and evolution of the trace through time provides almost identical results if a 10⁵-member ensemble is used and truncation of the SG method is made at 10. However, some small differences are still seen for both mean and standard deviations: for example, for the cloud droplets in Figure 2 or water vapor in Figure 3.

However, in NWP such large ensembles are not possible due to the high dimensionality of the problem that needs to be solved and the required computational time and costs. Therefore it is instructive to compare these two approaches once 50 ensemble members are used or SG is truncated at two. The 50 ensemble member statistics and statistics of SG truncated at two are not directly comparable via computational costs, but illustrate the situation of one taking a smaller than needed ensemble size (respectively smaller than needed truncation) in two methods. The results are illustrated in Figure 2 for the final time of the simulation at $t = 200$ s. In this case, the differences in standard deviations become significant in the two approaches. Figure 2 reveals that, in this case, using 50 ensemble members provides larger areas of underestimated uncertainties compared with the simulation with 10⁵ members, while comparing standard deviations of $M = 10$ and $M = 2$ simulations suggests that SG underestimates or overestimates the uncertainty in certain areas. The time evolution in Figure 3 shows large deviations in the normalized trace (gray lines) between 10⁵ and 50 simulations, depending on which 50 of the 10⁵ members are chosen, indicating that both over- and underestimation of the total trace is possible.

The differences in performance of the two methods depend on the system state. This dependence is shown in Figure 4 for three time instances, through a comparison of the error of the mean and error of the standard deviations, where the error is defined as the difference from 10⁵ members for MC and the difference from $M = 10$ for SG. Early in the simulation (at time $t = 20$ s), in more stable conditions, the SG method outperforms an ensemble of simulations for every truncation and every variable. In unstable conditions (at the end of the simulation) and in the presence of turbulence, an ensemble of simulations would need more than 100 members (depending on the model variable) and the SG method more than a truncation at five, to produce results below the gray line in Figure 4 that depicts the difference between mean and standard deviations calculated from 10⁵ ensembles and SG with $M = 10$. Note that, as time progresses, this difference increases as well. At the end of the simulation, the increase

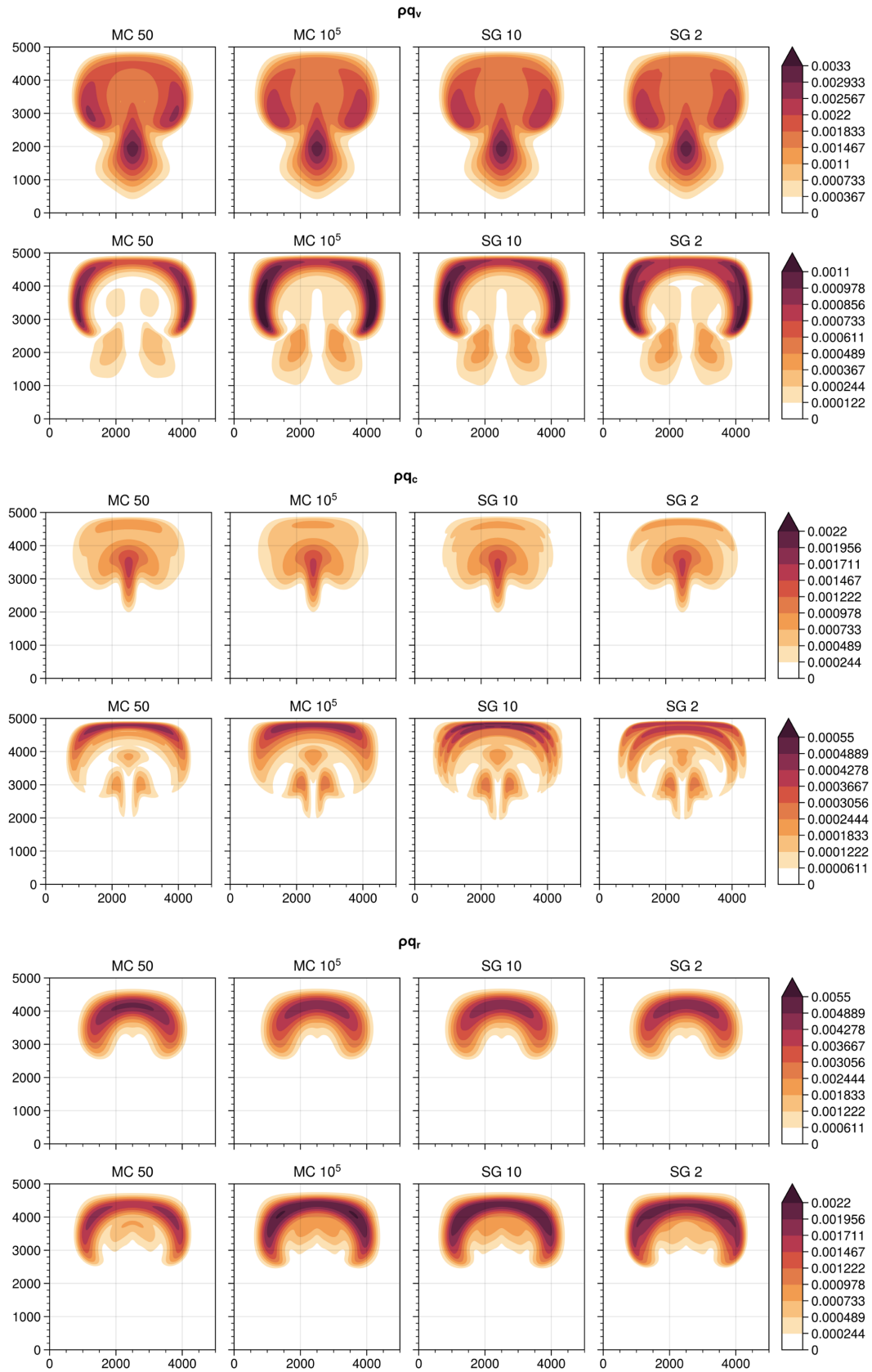


FIGURE 2 Mean (top) and standard deviations (bottom) of the mass densities of (a) water vapor ρ_{qv} , (b) cloud drops ρ_{qc} , and (c) rain ρ_{qr} , at time $t = 200$ s simulated with the MC method (50 and 10^5 members) and the SG method (two and 10 basis polynomials). [Colour figure can be viewed at wileyonlinelibrary.com]

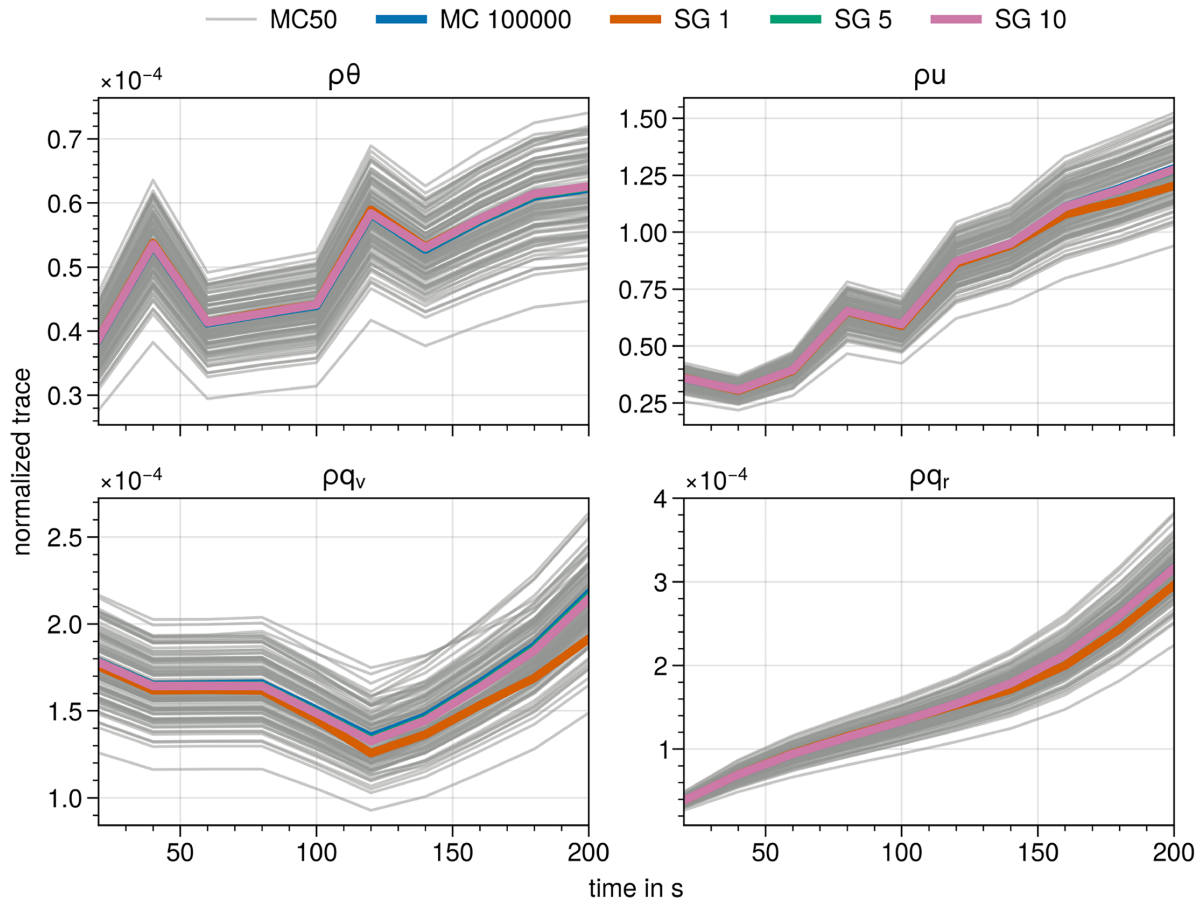


FIGURE 3 Evolution of the domain-averaged standard deviation (i.e., normalized trace of the covariance matrix) of the potential temperature $\rho\theta$, the horizontal velocity ρu , and the mass densities of water vapor ρq_v and cloud drops ρq_r simulated with the 10^5 -member ensemble and the SG method for $M = 1, 5$, and 10 . Note that the green line for $M = 5$ is not seen on the plot, since the pink line ($M = 10$) is on top of it. Gray lines indicate traces of 50 members chosen randomly from the 10^5 -member ensemble. [Colour figure can be viewed at wileyonlinelibrary.com]

is especially strong (order of magnitude) for u and q_v , indicating that for some variables in the case of unstable and turbulent flow we would need even more than $M = 10$ modes to represent the true solution with SG, and possibly even more than 10^5 ensemble members.

5 | DATA ASSIMILATION

In this section, we investigate the value of an SG uncertainty quantification method for data assimilation. To this end, we perform twin experiments. In an identical twin configuration, a model run is chosen as the nature run, to describe the evolution of the true state \mathbf{w} , consisting of density, velocity vector, moist potential temperature and concentration of water vapor, cloud droplets, and rain drops at each grid point of the model. Synthetic observations \mathbf{w}^{obs} are created from the nature run by adding observation errors e^{obs} as random perturbations. We chose to observe the q_r field every 20 s and at every 50th grid

point in each direction. Observations contain Gaussian observation noise with zero mean and standard deviation of 0.5×10^{-5} . The observation-error covariance matrix \mathbf{R} is taken to be diagonal, with values on the diagonal corresponding to the variance of the distributions used for generating the observation-error vector e^{obs} . The observation operator \mathbf{H} in our case is a matrix with elements zero or one, which determines the location and physical nature of the observations.

The stochastic EnKF (Evensen, 1994; Evensen, 2003) will be used for data assimilation in our experiments. The EnKF employs an ensemble of predictions (backgrounds) $\{\mathbf{w}^{\text{b},i}\}_{i=1}^N$ to calculate the sample background-error covariance matrix \mathbf{P}^{b} :

$$\mathbf{P}^{\text{b}} = \frac{1}{N-1} \sum_{i=1}^N [\mathbf{w}^{\text{b},i} - \bar{\mathbf{w}}^{\text{b}}] [\mathbf{w}^{\text{b},i} - \bar{\mathbf{w}}^{\text{b}}]^{\text{T}}, \quad (12)$$

with $\bar{\mathbf{w}}^{\text{b}}$ representing the ensemble mean, and N denoting the number of ensemble members. At the times for which

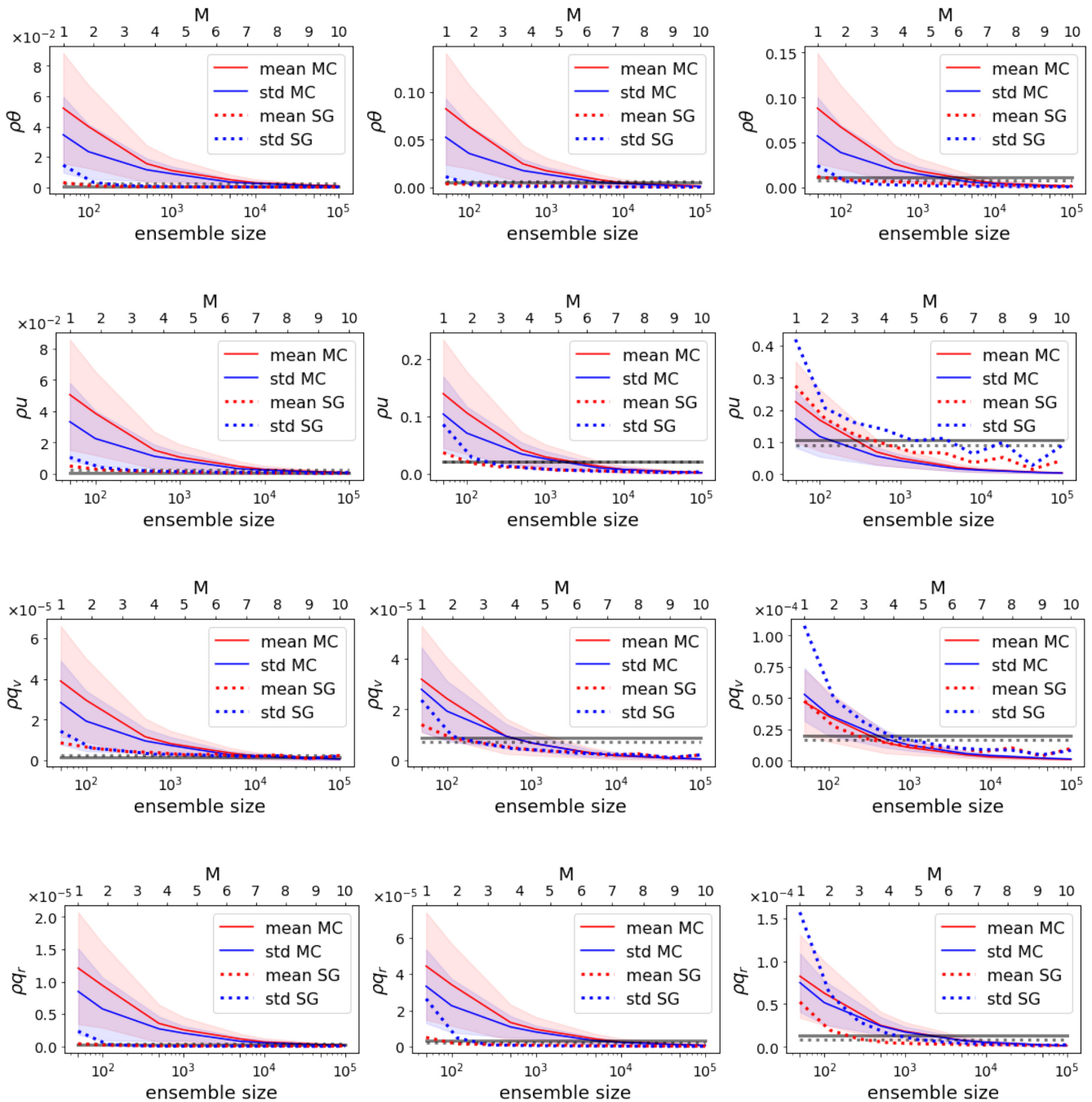


FIGURE 4 Comparison at $t = 20$ s (left), $t = 120$ s (middle), and $t = 200$ s (right) of the error of the mean (red) and standard deviation (blue) obtained with MC (solid) as a function of ensemble size and SG (dotted) as a function of M for density-weighted potential temperature, horizontal velocity, and the mass densities of water vapor ρq_v and cloud drops ρq_r . The error is defined as the difference from 10^5 members for MC and the difference from $M = 10$ for SG. Note that the x -axis is on a log scale for MC and linear for SG. For each ensemble size of MC, experiments were repeated 100 times. The standard deviations of these 100 experiments are shown through red (mean) and blue (std) shading. Gray lines depict the difference between mean (solid) and standard deviations (dotted) calculated from 10^5 ensembles and SG with $M = 10$. [Colour figure can be viewed at wileyonlinelibrary.com]

observations are available, each background ensemble is updated based on observations using the Kalman filter formula, as

$$\mathbf{w}^{a,i} = \mathbf{w}^{b,i} + \mathbf{P}^b \mathbf{H}^T (\mathbf{H} \mathbf{P}^b \mathbf{H}^T + \mathbf{R})^{-1} (\mathbf{w}^{obs,i} - \mathbf{H} \mathbf{w}^{b,i}), \quad (13)$$

where $\mathbf{w}^{obs,i} = \mathbf{w}^{obs} + \epsilon^{obs,i}$ are perturbed observations. Note that the sample covariance \mathbf{P}^b is singular when N is smaller than the size of the state vector, and due to the sampling error many of its elements will be inaccurate. This control experiment based on the MC method will be

compared with the EnKF modified for use with the SG method, as described next.

5.1 | EnKF with SG (SGEnKF)

We follow Li and Xiu (2009) and calculate the background-error covariance directly from the state vector of the SG coefficients $\hat{\mathbf{w}}$ as

$$\mathbf{P}^{\text{b,SG}} = \sum_{k=1}^M k! [\hat{\mathbf{w}}_k (\hat{\mathbf{w}}_k)^\top],$$

and, with formula (13), produce an analysis ensemble in the physical space. Note that, since the ensemble $\{\mathbf{w}^{\text{b},i}\}_{i=1}^{N_{\text{SG}}}$ is not evolved in time in this approach, we can generate many members cheaply, using

$$\mathbf{w}^{\text{b},i} = \sum_{k=0}^M \hat{\mathbf{w}}_k \Phi_k(\omega^i), \quad i = 1, \dots, N_{\text{SG}}.$$

The analysis ensemble $\{\mathbf{w}^{\text{a,SG},i}\}_{i=1}^{N_{\text{SG}}}$ from Equation (13) is used for obtaining M SG coefficients by using the orthogonality property (5) and evaluating the expected value:

$$\hat{\mathbf{w}}_k = \frac{1}{k!} \mathbb{E}[\mathbf{w}^{\text{a,SG}} \Phi_k(\omega)] = \int_{\mathbb{R}} \mathbf{w}^{\text{a,SG}} \Phi_k(\omega) \mu(\omega) d\omega. \quad (14)$$

The expected value can be approximated either by an averaged discrete sum

$$\frac{1}{k! N_{\text{SG}}} \sum_{i=1}^{N_{\text{SG}}} \mathbf{w}^{\text{a,SG},i} \Phi_k(\omega^i),$$

where ω^i are drawn randomly, or, since ω is normally distributed, the expectation could be calculated using Gauss–Hermite quadrature, where ω^i are quadrature points. We find the second approach to be less expensive and use it in our experiments with 100 quadrature points. Comparable results with $M = 3$ for the first approach would require us to generate $N_{\text{SG}} = 10\,000$ members. We also note that, if M increases, more synthetic members would be needed in the first approach to obtain all of the SG coefficients correctly, since their norms decay exponentially with M .

5.2 | Accuracy of analysis and short term forecasts

The EnKF with SG (SGEnKF) with $M = 3$ has already produced accurate results, as illustrated with the

root-mean-square error (RMSE) in Figure 5 (left) for forecasts of u and q_r , although the RMSE–spread ratio for u with $M = 6$ is better than for $M = 3$. We attribute the already good performance of the SGEnKF with $M = 3$ to the ability of SG with the choice of normally distributed ω to represent the analysis uncertainty well (this is calculated using the Kalman filter equations and therefore by design steered toward a Gaussian distribution). The RMSE and RMSE–spread ratio for the EnKF with $N_{\text{ens}} = 10, 50$ and the SGEnKF with $M = 6$ are shown in Figure 5 (right). The SGEnKF with $M = 3$ and $M = 6$ shows higher accuracy than the EnKF with 10 or 50 members. However, when reducing M to 2, the accuracy deteriorates. Figure 4 and the divergence of data assimilation results for two random modes shown in Figure 5 indicate that such a small number of modes is not sufficient to describe the evolution of the covariances, as the dynamics starts to be more complex and vortex structures arise.

In Figure 6, the RMSE and RMSE–spread ratio of the EnKF with $N_{\text{ens}} = 10, 50$ and the SGEnKF with $M = 6$ are shown as a function of the initial accuracy of the q_v field. Although the initial uncertainty of q_v is fixed to 20%, the initial error depends on the realization of the truth and, in the case of the EnKF, on the realization of the initial ensemble. For example, high initial error can result in a scenario of an unusual weather situation, that is, a true field being a realization in tails of the distribution. If the initial error is high, the RMSE and RMSE–spread ratios at the end of the data assimilation experiment are higher as well, for both methods. However, the EnKF is much more sensitive to the initial error. Repeating the experiment for various initial RMSEs shows that the RMSE of the EnKF with different initial realizations, even with 50 members, is higher than that of the SGEnKF with $M = 6$, especially for higher initial error. Therefore, the SGEnKF seems to be more robust than the EnKF with respect to realization of the truth.

Although the EnKF is used with both uncertainty quantification methods, differences in the data assimilation algorithm exist. For one, ensemble methods require localization in practice, to increase the rank of covariances (Janjić *et al.*, 2011), while the SGEnKF can generate many synthetic members, which are needed for accurate reconstruction of the SG analysis coefficients. Further, differences between these two methods can appear in the case of nonlinear observation operators. While, for the SGEnKF, nonlinear observation operators can be used on each $\mathbf{w}^{\text{b},i}$ in Equation (13), calculation of the analysis with $\mathbf{P}^{\text{b,SG}}$ would require linearization of the nonlinear observation operator for approximating matrices $\mathbf{P}^{\text{b}}\mathbf{H}^\top$ and $\mathbf{H}\mathbf{P}^{\text{b}}\mathbf{H}^\top$. An alternative approach would be to use N_{SG} members and calculate these covariances following Equation (12), which might be costly, depending on the

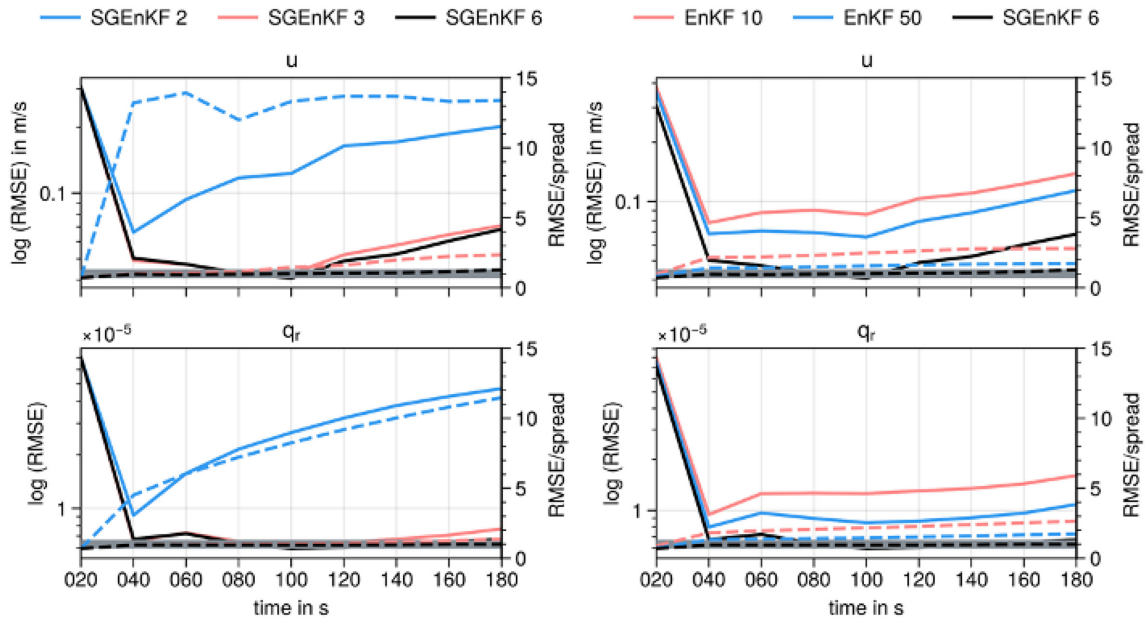


FIGURE 5 Evolution of the log of forecast root-mean-square error (RMSE) and RMSE–spread ratio (dashed lines) for SGenKF with $M = 2, 3,$ and 6 (left) and for EnKF with $N_{ens} = 10, 50$ and SGenKF with $M = 6$ (right). Results are shown for variables u (upper) and q_r (lower). Logs of RMSE are plotted as solid lines (see legend), with the gray tick line representing a perfect RMSE–spread ratio of one. [Colour figure can be viewed at wileyonlinelibrary.com]

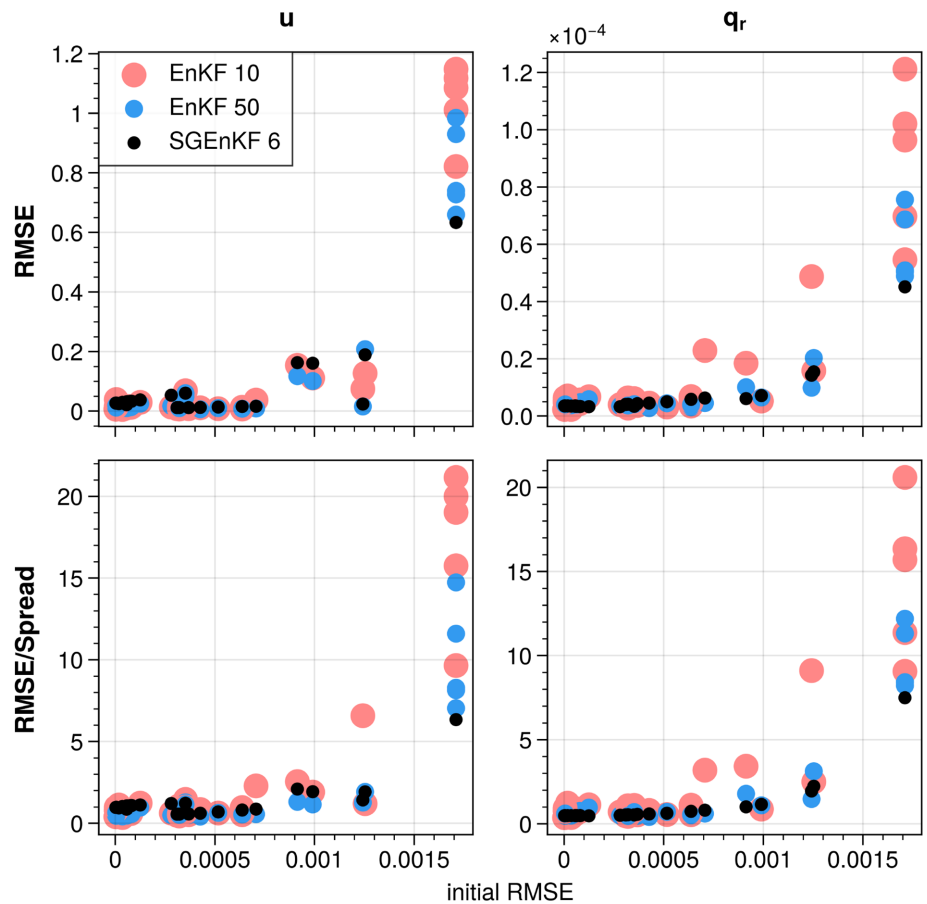


FIGURE 6 RMSE (upper) and RMSE–spread ratio (lower) for u (left) and q_r (right) at the end of the data assimilation experiment depending on initial accuracy in the q_r field. [Colour figure can be viewed at wileyonlinelibrary.com]

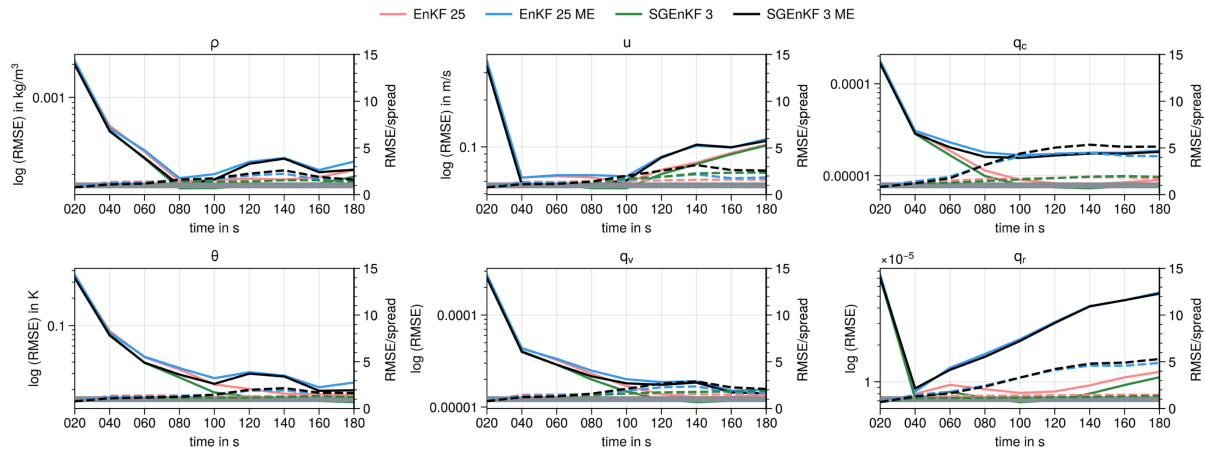


FIGURE 7 Evolution of the log of forecast RMSE and RMSE–spread ratio (dashed lines) for SGEnKF with $M = 3$ and for EnKF with $N_{\text{ens}} = 25$. The case in which the true field contains the model error is marked with ME. Results are shown for variables ρ , u , q_c (upper) and θ , q_v , q_r (lower). Logs of RMSEs are plotted as solid lines (see legend), with a gray tick line representing a perfect RMSE–spread ratio of one. [Colour figure can be viewed at wileyonlinelibrary.com]

number of observations. In this case, the computational cost of the method increases. However, the SGEnKF would require a model evolution for only $M + 1$ coefficients, which is much smaller than the N ensemble members needed for the EnKF. Note that the model evolution is usually the most expensive component of the data assimilation system.

Finally, we compare the two methods in the case in which the truth is generated with a parameter value that is different from the one used in the numerical model, mimicking the case of the presence of model error. We chose to perturb the parameter that influences the terminal velocity of a droplet (Chertock *et al.*, 2019). In Chertock *et al.* (2019), the sensitivity of the cloud variables to this parameter (parameter α in their study) was investigated. To generate the truth field for the data assimilation experiments, a 10% random perturbation was added to the parameter. The setup is otherwise the same as for the perfect model experiments, so only q_r is observed. Figure 7 illustrates the results for forecasts of ρ , u , q_c , θ , q_v , q_r , for the SGEnKF with $M = 3$ and the EnKF with 25 members, in cases with and without model error. In the presence of model error, the behavior of both methods is very similar for all model variables, indicating that one algorithm is not more sensitive to the presence of model error than the other. Although not done in our experiment, for both methods the same techniques could be used to take into account model error due to the unknown parameter during data assimilation.

6 | CONCLUSIONS

In order potentially to use the SG method (Chertock *et al.*, 2023) for the evolution of uncertainty through time

in atmospheric applications, we performed a first comparison study between the ensemble approach and the SG method on an idealized two-dimensional model that couples nonhydrostatic weakly compressible Navier–Stokes equations to cloud variables. The two methods are compared for probabilistic forecasting and in data assimilation settings. In this first step, the numerical model is chosen to be simple, to be able to compute ensemble statistics accurately. Further, in this article the uncertainty arises only as a parameter in the initial data; however, the SG method presented also applies directly to other uncertain parameters in the model, for example in the cloud parametrizations. For efficiency reasons, we have only considered one-dimensional uncertainty. However, the same strategy as described for the forecast uncertainty, as well as the data assimilation, can be used when moderate (approximately 10) independent uncertainties are considered. In such a case we end up with a multidimensional stochastic space and ω will be a vector of independent random variables. Despite the clarity of such an approach, there is a limitation in its realization, since the high-dimensional stochastic space may yield a curse of dimensionality: see, for example, (Doostan and Iaccarino, 2008). One way to overcome these restrictions is the use of a special sparse approximation and adaptive approximation in stochastic space. Nevertheless, for a moderate number of uncertain data/parameters, the SG method presented in this article is an attractive alternative for forecast uncertainty as well as for data assimilation.

A series of experiments indicates that differences in the performance of the two methods depend on the system state and truncations used. Once approximations are applied, for example with 50 ensemble members, or the SG expansion is truncated at two, differences in standard deviations become significant in the two approaches. The

differences in performance of the two methods depend on the system state as well. For example, in more stable conditions, the SG method outperforms the ensemble of simulations for every truncation and every variable. In unstable conditions, the ensemble of simulations would need more than 100 members (depending on the model variable) and the SG method more than five modes to produce comparable but not identical results. Further, in unstable conditions the uncertainty as represented with mean and standard deviation between the 10^5 -member ensemble and the truncation of the SG method at 10 basis polynomials still exists.

We also used the SG method instead of an ensemble of simulations within data assimilation. Our experiments concentrated on the ability of both methods to correct unobserved variables through their representation of covariances, therefore we only observed q_r . Forecast RMSEs resulting from the EnKF when increasing the number of ensemble members converged towards the RMSEs resulting from the SGenKF. The SGenKF with $M = 3$ already produced accurate results. This good performance with a nonlinear model is probably due to both the ability of SG to represent analysis uncertainty well with a low number of modes and its exponential convergence rates. Note that, since we used Kalman filter equations to calculate the analysis uncertainty, which by design is Gaussian-distributed, it was already possible for SG to represent this initial uncertainty well with a low number of modes. Although RMSEs with $M = 3$ with the SGenKF were already small, improvement was still possible in the RMSE–spread ratio for all variables by increasing to $M = 6$. We also showed that SGenKF is less sensitive in a perfect model scenario to possible extreme realizations of the true field. In the presence of model error, in our case due to an unknown parameter, the SGenKF and the EnKF produced very similar results. Additive inflation as well as other algorithms used in EnKF settings to treat model error could additionally be applied to the SGenKF if needed. It is also important to note that, for data assimilation, nothing limits the use of the SG method with other sample methods more appropriate for the highly non-Gaussian problem found in this application, such as the quadratic programming ensemble (Janjić *et al.*, 2014), the quadratic filter (Hodyss, 2012), or particle filters (van Leeuwen, 2009), rather than the EnKF as in this study.

AUTHOR CONTRIBUTIONS

T. Janjić: conceptualization; funding acquisition; methodology; supervision; writing – original draft. **M. Lukáčová-Medviďová:** conceptualization; funding acquisition; methodology; supervision; writing – review and editing. **Y. Ruckstuhl:** formal analysis;

investigation; methodology; software; validation; visualization; writing – review and editing. **B. Wiebe:** formal analysis; investigation; methodology; software; validation; visualization; writing – original draft.

ACKNOWLEDGEMENTS

The research leading to these results has been performed within subprojects A2 and B6 of the Transregional Collaborative Research Center SFB/TRR 165 “Waves to Weather” (Craig *et al.*, 2021), funded by the German Science Foundation (DFG). Further, Tijana Janjić thanks the DFG for her Heisenberg Program Award (JA 1077/4-1), Mainz Institut of Multiscale Modelling: M3odel, and Gutenberg Research College: JGUMainz. Open Access funding enabled and organized by Projekt DEAL.

ORCID

T. Janjić  <https://orcid.org/0000-0002-8837-0879>

M. Lukáčová-Medviďová  <https://orcid.org/0000-0002-4351-0161>

REFERENCES

- Bryan, G.H. and Fritsch, J.M. (2002) A benchmark simulation for moist nonhydrostatic numerical models. *Monthly Weather Review*, 130, 2917–2928.
- Chertock, A., Kurganov, A., Lukáčová-Medviďová, M., Spichtinger, P. and Wiebe, B. (2019) Stochastic Galerkin method for cloud simulation. *Mathematics of Climate and Weather Forecasting*, 5(1), 65–106.
- Chertock, A., Kurganov, A., Lukáčová-Medviďová, M. and Wiebe, B. (2023) Stochastic Galerkin method for cloud simulation. Part II: a fully random Navier-stokes-cloud model. *Journal of Computational Physics*, 479, 111987.
- Craig, G.C., Fink, A.H., Hoese, C., Janjić, T., Knippertz, P., Laurian, A., Lerch, S., Mayer, B., Miltenberger, A., Redl, R., Riemer, M., Tempest, K.I., Wirth, V. (2021) Waves to weather: Exploring the limits of predictability of weather. *Bulletin of the American Meteorological Society*, 102, E2151–E2164. <https://doi.org/10.1175/BAMS-D-20-0035.1>.
- Davies, R.M. and Taylor, F.J. (1950) The mechanism of large bubbles rising through extended liquids and through liquids in tubes. *Proceedings of the Royal Society London A*, 200, 375–390.
- Després, B., Poëtte, G. and Lucor, D. (2013) Robust uncertainty propagation in systems of conservation laws with the entropy closure method. In: *Uncertainty Quantification in Computational Fluid Dynamics*. Heidelberg: Springer, pp. 105–149.
- Doostan, A. and Iaccarino, G. (2008) Breaking the curse of dimensionality for a class of pdes with stochastic inputs. In: *Annual Research Briefs*. Stanford, CA: Center of Turbulence, pp. 19–30.
- Elman, H.C., Miller, C.W., Phipps, E.T. and Tuminaro, R.S. (2011) Assessment of collocation and Galerkin approaches to linear diffusion equations with random data. *International Journal for Uncertainty Quantification*, 1(1), 19–33.
- Evensen, G. (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, 99, 10143–10162.

- Evensen, G. (2003) The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53, 343–367.
- Feireisl, E. and Lukáčová-Medviďová, M. (2023) Convergence of a stochastic collocation finite volume method for the compressible Navier-stokes system. *Annals of Applied Probability*. <https://doi.org/10.48550/arXiv.2111.07435>
- Feireisl, E., Lukáčová-Medviďová, M., She, B. and Yuan, Y. (2022) Convergence and error analysis of compressible fluid flows with random data: Monte Carlo method. *Mathematical Models and Methods in Applied Sciences*, 32(14), 2887–2925.
- Gustafsson, N., Janjić, T., Schraff, C., Leuenberger, D., Weissman, M., Reich, H., Brousseau, P., Montmerle, T., Wattrelot, E., Buaneck, A., Mile, M., Hamdi, R., Lindsog, M., Barkmeijer, J., Dahlbom, M., Macpherson, B., Ballard, S., Inverarity, G., Carley, J., Alexander, C., Dowell, D., Liu, S., Ikuta, Y. and Fujita, T. (2018) Survey of data assimilation methods for convective-scale numerical weather prediction at operational centres. *Quarterly Journal of the Royal Meteorological Society*, 144(713), 1218–1256.
- Hodyss, D. (2012) Accounting for skewness in ensemble data assimilation. *Monthly Weather Review*, 140, 2346–2358.
- Janjić, T., McLaughlin, D., Cohn, S.E. and Verlaan, M. (2014) Conservation of mass and preservation of positivity with ensemble-type Kalman filter algorithms. *Monthly Weather Review*, 142(2), 755–773.
- Janjić, T., Nerger, L., Albertella, A., Schroeter, J. and Skachko, S. (2011) On domain localization in ensemble based Kalman filter algorithms. *Monthly Weather Review*, 139, 2046–2060. <https://doi.org/10.1175/2011MWR3552.1>.
- Janjić, T., Ruckstuhl, Y. and Toint, P.L. (2021) A data assimilation algorithm for predicting rain. *Quarterly Journal of the Royal Meteorological Society*, 147(736), 1949–1963.
- Li, J. and Xiu, D. (2009) A generalized polynomial chaos based ensemble kalman filter with high accuracy. *Journal of Computational Physics*, 228(15), 5454–5469.
- Loh, W.-L. (1996) On Latin hypercube sampling. *Annals of Statistics*, 24(5), 2058–2080.
- Lukáčová-Medviďová, M. and Schömer, A. (2022) Existence of dissipative solutions to the compressible Navier-stokes system with potential temperature transport. *Journal of Mathematical Fluid Mechanics*, 24(3), 82. <https://doi.org/10.1007/s00021-022-00713-3>
- Lukáčová-Medviďová, M. and Schömer, A. (2023) Compressible Navier-stokes equations with potential temperature transport: stability of the strong solution and numerical error estimates. *Journal of Mathematical Fluid Mechanics*, 25(1), 1–38.
- Ma, X. and Zabarar, N. (2009) An adaptive hierarchical sparse grid collocation algorithm for the solution of stochastic differential equations. *Journal of Computational Physics*, 228(8), 3084–3113.
- Mishra, S. and Schwab, C. (2012) Sparse tensor multi-level Monte Carlo finite volume methods for hyperbolic conservation laws with random initial data. *Mathematics of Computation*, 81(280), 1979–2018.
- Mishra, S., Schwab, C. and Šukys, J. (2013) Multi-level Monte Carlo finite volume methods for uncertainty quantification in nonlinear systems of balance laws. In: *Uncertainty Quantification in Computational Fluid Dynamics*, vol. 92 of Lect. Notes Comput. Sci. Eng. Heidelberg: Springer, pp. 225–294.
- Morokoff, W.J. and Caflisch, R.E. (1995) Quasi-Monte Carlo integration. *Journal of Computational Physics*, 122(2), 218–230.
- Niederreiter, H. (1992) *Random number generation and quasi-Monte Carlo methods*, vol. 63 of CBMS-NSF Regional Conference Series in Applied Mathematics. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).
- Palmer, T. (2019) The ecmwf ensemble prediction system: looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society*, 145(1), 12–24.
- Poëtte, G., Després, B. and Lucor, D. (2009) Uncertainty quantification for systems of conservation laws. *Journal of Computational Physics*, 228(7), 2443–2467.
- Richardson, D.S. (2000) Skill and relative economic value of the ecmwf ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 126(563), 649–667.
- Ruckstuhl, Y. and Janjić, T. (2018) Parameter and state estimation with ensemble Kalman filter based approaches for convective scale data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 144(712), 826–841.
- Stein, M. (1987) Large sample properties of simulations using Latin hypercube sampling. *Technometrics*, 29(2), 143–151.
- Šukys, J., Mishra, S. and Schwab, C. (2013) *Multi-level Monte Carlo finite difference and finite volume methods for stochastic linear hyperbolic systems*. *Monte Carlo and Quasi-Monte Carlo Methods 2012*, vol. 65 of Springer Proc. Math. Stat. Heidelberg: Springer, pp. 649–666.
- Tryoen, J., Le Maître, O., Ndjinga, M. and Ern, A. (2010) Intrinsic Galerkin methods with upwinding for uncertain nonlinear hyperbolic systems. *Journal of Computational Physics*, 229(18), 6485–6511.
- van Leeuwen, P.J. (2009) Particle filtering in geophysical systems. *Monthly Weather Review*, 137(12), 4089–4114.
- Wan, X. and Karniadakis, G.E. (2006) Long-term behavior of polynomial chaos in stochastic flow simulations. *Computer Methods in Applied Mechanics and Engineering*, 195(41-43), 5582–5596.
- Wiebe, B. (2021) *Numerical Simulations and Uncertainty Quantification for Cloud Simulation*. PhD thesis. Mainz: Johannes Gutenberg-University.
- Witteveen, J.A.S., Loeven, A. and Bijl, H. (2009) An adaptive stochastic finite elements approach based on Newton-cotes quadrature in simplex elements. *Computers & Fluids*, 38(6), 1270–1288.
- Xiu, D. and Hesthaven, J.S. (2005) High-order collocation methods for differential equations with random inputs. *SIAM Journal on Scientific Computing*, 27(3), 1118–1139.
- Xiu, D. and Karniadakis, G.E. (2002a) The wiener-Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2), 619–644.
- Xiu, D. and Karniadakis, G.E. (2002b) The wiener-Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2), 619–644.

How to cite this article: Janjić, T., Lukáčová-Medviďová, M., Ruckstuhl, Y. & Wiebe, B. (2023) Comparison of uncertainty quantification methods for cloud simulation. *Quarterly Journal of the Royal Meteorological Society*, 149(756), 2895–2910. Available from: <https://doi.org/10.1002/qj.4537>