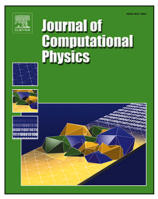
Probabilistic data-driven turbulence closure modeling by assimilating statistics

Journal of Computational Physics 539 (2025) 114234

Contents lists available at ScienceDirect

Journal of Computational Physics

journal homepage: www.elsevier.com/locate/jcp



Probabilistic data-driven turbulence closure modeling by assimilating statistics

Sagy R. Ephrati *

Department of Mathematical Sciences, Chalmers University of Technology, University of Gothenburg, 412 96, Gothenburg, Sweden

ARTICLE INFO

Keywords:
Sub-grid scale modeling
Turbulence
Stochastic
Bayesian
Data-driven
Data assimilation

ABSTRACT

A framework for deriving probabilistic data-driven closure models is proposed for coarse-grained numerical simulations of turbulence in statistically stationary state. The approach unites the *ideal large-eddy simulation* model [8] and data assimilation methods. The method requires *a posteriori* measured data to define a stochastic large-eddy simulation model, which is combined with a Bayesian statistical correction enforcing user-specified statistics extracted from high-fidelity flow snapshots. Thus, it enables computationally cheap ensemble simulations by combining knowledge of the local integration error and knowledge of desired flow statistics. An example implementation of the modeling framework is given for two-dimensional Rayleigh-Bénard convection at Rayleigh number $Ra = 10^{10}$, incorporating stochastic perturbations and an ensemble Kalman filtering step in a non-intrusive way. Physical flow dynamics are obtained, whilst kinetic energy spectra and heat flux are accurately reproduced in long-time ensemble forecasts on coarse grids for two discretizations. The model is shown to produce accurate results with as few as 20 high-fidelity flow snapshots as input data.

1. Introduction

The highly turbulent nature of fluid flows provides a major challenge in the prediction of fluid-dynamical processes. Examples include oceanic and atmospheric flows, relevant to geophysics and climate science, or engineering and industrial applications involving thermal processes. The nonlinearity in the governing equations causes interaction between flow features of different scales, leading to energy distribution over a wide range of scales of motion [1]. Accurate direct numerical simulations (DNS) of fluid-dynamical models thus require very fine computational grids to fully resolve all flow features, and quickly become prohibitively expensive to carry out. Feasible simulation strategies for such systems therefore necessitate a reduction in computational complexity to approximate the flow evolution. Well-established approaches dealing with complexity reduction include reduced-order modeling [2], where the governing equations or associated operators are formulated in a low-dimensional manner, and large-eddy simulation (LES) [3,4], where the governing equations are spatially filtered and small-scale effects are modelled. In the current work, we focus on the LES approach and combine this with stochastic forecasting and data assimilation. The result is a general modeling framework for data-driven error correction in low-fidelity fluid predictions, which we demonstrate for two-dimensional Rayleigh-Bénard convection. We elaborate on the various model components below.

Large-eddy simulation. The spatially filtered equations of motion form the starting point for large-scale LES models. The underlying rationale is that computationally feasible methods can be developed by only capturing the largest scales of motion, since these are most

* Corresponding author.
E-mail address: sagy@chalmers.se

https://doi.org/10.1016/j.jcp.2025.114234
Received 25 March 2025; Received in revised form 18 June 2025; Accepted 8 July 2025
Available online 12 July 2025

important for the overall flow evolution. The level of detail in the LES solution is determined by the filter width. Explicit LES follows from applying the adopted spatial filter to the governing equations directly. Alternatively, the equations can be filtered implicitly by discretizing differential operators on coarse computational grids. Regardless of the chosen approach, filtering introduces errors and uncertainty in the flow evolution. The filter generally does not commute with the nonlinear terms in the governing equations, yielding a so-called commutation error. In addition, discretization error is introduced on coarse computational grids due to poorly resolved spatial derivatives. One can add modeling terms to the governing equations with the aim of compensating for the errors introduced by filtering and coarsening [5–7]. We will refer to this as the LES model, although it is also often referred to as the closure model or the sub-filter or sub-grid scale model. Furthermore, filtering introduces uncertainty since it discards information of small spatial scales. Infinitely many fully resolved flow fields will correspond to the same filtered LES field, which makes it impossible to know the exact evolution of the LES solution. Consequently, it is desirable to develop error-correcting LES models that simultaneously quantify the uncertainty in the LES solution.

The abstract ideal LES model was derived by Langford and Moser [8] and can serve as a guideline for model development. The model is ideal in the sense that it reproduces single-time, multi-point statistics of the unfiltered solution and minimizes error in the instantaneous dynamics. It states that the ideal evolution of an LES solution is the average evolution conditional to an infinite number of fully resolved (unfiltered) fields corresponding to the current large-scale (filtered) field. One can probe this distribution of fields using a stochastic forecast ensemble and subsequently approximate the conditional average. Naturally, the model depends on the governing equations, the adopted filter and, in simulations, on the chosen spatial and temporal discretizations. This makes it challenging to approximate the conditional average well. However, empirical approximations can still be found through the use of data.

The field of data-driven LES has seen rapid development in recent years due to the increasing availability of computational resources and accessibility of high-fidelity data. Machine learning is used regularly to train neural networks as LES models [9,10]. One can distinguish between *a priori* and *a posteriori* learning, which differ in their objective function. *A priori* methods rely only on the high-fidelity data to learn a closure model, for example by minimizing a function of the unfiltered and the filtered solution. This approach can yield accurate short-time predictions, but might also suffer from errors and instabilities due to model-data inconsistency [11,12]. *A posteriori* methods can overcome this inconsistency, by directly comparing the results of the low-fidelity solver and the high-fidelity data [13]. Consequently, the model will depend on the configuration of the low-fidelity solver. The presently proposed model uses an *a posteriori* approach based on the ideal LES formalism to calibrate a stochastic subgrid-scale model.

Stochastic modeling in coarsened fluid descriptions. Stochasticity is commonly used in geophysical fluid dynamics to model uncertainty [14,15] stemming from imperfect initial conditions and incomplete models [16]. A prevalent interpretation of stochastic forcing is that it models the chaotic behavior of unresolved small scales on the resolved dynamics. For example, data-informed stochastic models have been developed for the two-scale Lorenz '96 system [17]. This system serves as a low-dimensional model of the atmosphere with variables evolving over different time scales. By replacing the influence of fast variables on the slow variables with data-driven stochastic processes, [18] were able to account for model error. The addition of time-correlated noise further improved the forecasting skill. Other approaches based on data-driven Markov chains [19] have led to state-dependent stochastic processes modeling the effects of unresolved variables yielding a good reproduction of statistics of the resolved variables, such as probability density functions and temporal autocorrelation. Alternatively, [20] produce an ensemble of predictions in latent space via data-driven neural stochastic differential equations and subsequently average the corresponding LES predictions to approximate the ideal LES prediction. They report reduced errors over long simulation times and good predictions of the time-averaged energy spectrum when compared to implicit LES and deterministic neural network closures.

The current method fits in the recent modeling trend in observational sciences where high resolution simulations are replaced by stochastically forced simulations on coarse grids. New types of data-driven stochastic models have been developed for fluid-dynamical systems for the purpose of modeling uncertainty in observations and correct for undesired coarsening effects. For example, the effect of transport noise [21,22] in advection-dominated geophysical flows has been studied in the context of uncertainty quantification [23–25] and data assimilation [26]. These studies showed good quantification of uncertainty in advective processes at severely reduced computational costs. Another advantage of stochastic forcing is that it allows for indefinite simulation of signals with desired statistical properties, for example mimicking features of the dataset from which the model is calibrated. The chaotic evolution of turbulent flows prompts the development of stochastic models that reproduce flow statistics rather than pointwise agreement with some reference. Inspired by the nonlocality of turbulence, a space- and time-dependent model can be decomposed via global basis functions for which only the time series are modeled [27,28], possibly based on statistical data. Examples include the use of include proper orthogonal functions (POD) [29] and Fourier modes to reproduce their respective spectra in coarse numerical simulations of Rayleigh-Bénard convection [30,31]. [32] employed spherical harmonic functions to develop a stochastic subgrid-scale term with memory effects for barotropic flow, and found good agreement in the resulting energy spectra. Capturing flow statistics instead of the full flow also reduces the amount of data required to calibrate the model. Low-dimensional models with basis functions tailored to user-specified statistics can be derived [33] and used to develop stochastic models at severely reduced computational costs while accurately reproducing selected statistics, as was recently shown for the two-dimensional Navier-Stokes equations [34].

Data assimilation and error correction. Data assimilation (DA) deals with efficiently incorporating data in predictions and is widely employed in geophysical sciences. DA combines predictions of dynamical systems with observations optimally, taking into account the uncertainty in both these aspects [35,36]. A common way to include information into a prediction ensemble is via Bayes' theorem [36,37]. This has led to DA schemes for nonlinear problems such as the ensemble Kalman filter (EnKF) [38,39] and particle filters [40].

In the context of DA, filtering[1] of turbulent systems has also been carried out by representing interactions of resolved and unresolved scales as stochastic processes [41]. Recent work has provided a theoretical framework for filtering of dynamical systems by observing statistics only [42], with correcting model errors as a possible application. The observations are often dealt with sequentially in DA, resulting in 'on-the-fly' updates of predictions or parameters as new information becomes available [43]. This contrasts with data-driven LES, where model parameters are commonly determined before performing a numerical simulation. Nonetheless, we will show here that DA techniques can be adapted to yield turbulence closure models for coarsened discretized systems.

Coarsening a discretized system will lead to a change in the predicted dynamics. Such a change is not necessarily a cause for concern at sufficiently high resolution. However, severe coarsening induces strong errors and yields numerical solutions converging to a different statistically steady state, as may be observed when measuring statistics of the dynamical system at large lead times. These structural errors can be alleviated by steering predictions towards observations (or specific regions of the state space), thereby increasing their likelihood. This is referred to as *nudging* [44]. A similar concept of continous data assimilation (CDA) based on interpolated approximations of measurements has been developed for fluid flows [45]. CDA assimilates observations while the predictions are integrated in time via a term that nudges the prediction towards the observation. Theoretical convergence for downscaled solutions has been proved for the two-dimensional Navier-Stokes equations [46], accompanied by numerical results for this system [47] and two-dimensional Rayleigh-Bénard convection [48,49]. The CDA approach is similar to the continuous-time 3D-Var DA scheme [50,51]. Recently, a heuristic data-driven stochastic closure model was derived from the 3D-Var scheme, with the aim of reproducing reference energy spectra. This has been demonstrated for the two-dimensional Euler equations [52], Rayleigh-Bénard convection [30], and quasi-geostrophic flow on the sphere [53]. However, unrealistic dynamics may arise when nudging the solution too strongly [36] and we therefore endeavour to replace ad-hoc modeling assumptions by a structured method based on DA techniques. This leads to a modeling framework from which we derive a nudging approach for turbulent flows. One of the results in this paper is that the proposed model yields realistic instantaneous dynamics while reproducing selected flow statistics in long-time simulations, even when only a small amount of data is available.

Contributions and paper outline. In this paper, we propose a probabilistic data-driven LES closure modeling strategy with the aim of correcting for undesired coarsening effects on the dynamics whilst modeling the inherent uncertainty. The approach is inspired by the abstract *ideal LES model* [8] and consists of a stochastic ensemble LES prediction followed by a Bayesian correction. Specifically, stochastic LES yields a computationally cheap ensemble prediction using *a posteriori* measured data. Subsequently, a Bayesian correction updates selected key statistics (or quantities of interest, QoIs) based on high-fidelity data. Thus, we combine knowledge of the local errors of the low-fidelity solver with knowledge of long-time statistics of the high-fidelity reference result. The model permits indefinite computationally efficient ensemble simulations of the dynamical system. An example of this methodology is provided for two-dimensional Rayleigh-Bénard convection, where we focus on kinetic energy spectra and heat flux as key statistics.

The paper is structured as follows. In Section 2, we recapitulate the ideal LES model and extend this to discrete time to set the stage for stochastic LES modeling. In Section 3, we recall sequential data assimilation and formulate the general closure model by assimilating statistics. An example of a closure model derived via the proposed modeling framework is described for Rayleigh-Bénard convection in Section 4. Section 5 presents an assessment of both of both short-time and long-time performance in the cases of plenty historical data and sparse historical data, and also examines the regularization properties of the proposed method. A discussion on model generalization is provided in Section 6. The paper is concluded in Section 7.

2. Ideal large-eddy simulation

LES generally starts from a filtered description of the governing equations of motion. The abstract ideal LES model [8] was derived to compensate for unwanted effects caused by underresolving the turbulent flow. It is ideal in the sense that it minimizes the instantaneous error in the evolution of the large-scale dynamics. In this section we recall the ideal LES model as presented by Langford and Moser [8]. We summarize Ideal LES in Section 2.1 and subsequently extend the formulation to discrete time in Section 2.2. In Section 2.3 we motivate stochastic modeling on coarse computational grids using ideal LES.

Notation. We follow the notation of [8] and denote by u and v unfiltered turbulent fields. The space of unfiltered fields is denoted by $\mathcal{V}$. We let w be a resolvable large-scale field, referred to as an LES field, and the space of filtered fields is denoted by $\mathcal{W}$. Filtered variables are denoted by a tilde $\tilde{\cdot}$. Distributions are denoted by π, which will be distinguished via subscripts. Time levels are denoted by superscripts, e.g., u^n denotes a turbulent field at time t^n. Sets of fields are written in curly brackets $\{\cdot\}$, where we distinguish between sequences of snapshots and ensembles at specific times. Sequences of snapshots of a solution u are denoted by $\{u^j\}$. An ensemble at time t^n is denoted by $\{u_i^n\}$, where the i^{th} ensemble member is the field u_i^n. As will become clear from context, we also augment the notation of ensembles with subscripts f and a to respectively denote forecast and analysis ensembles.

[1] The term *filtering* is used both the fields of LES and data assimilation but carries a different meaning in each of the fields. In LES, filtering refers to spatial filtering: a method to determine the level of physical detail to keep in the LES solution. In data assimilation, filtering deals with the sequential updating of a probability distribution function of a random variable as new measurements become available.

2.1. Ideal LES in continuous time

We consider a field of interest u with an evolution given by $L(u)$, which contains a nonlinear advection term. The numerical method that integrates L will be referred to as the high-fidelity solver. The filtered evolution of u reads

$$\widetilde{\frac{\partial u}{\partial t}} = \widetilde{L(u)}, \tag{1}$$

which can be expressed as the evolution of the filtered field $\tilde{u}$ as

$$\frac{\partial \tilde{u}}{\partial t} = L(\tilde{u}) + M(u). \tag{2}$$

In the last equation, $M(u) = \widetilde{L(u)} - L(\tilde{u})$ is a model term that appears due to the filtering of the nonlinear advection term, which causes a commutation error. At this stage, M is exact, but evaluating M requires explicit knowledge of the unfiltered field and will instead be replaced by a model.

The ideal subgrid model is shown to minimize error in the instantaneous evolution of large scales and produce accurate spatial statistics [8]. The fundamental insight in the model derivation is that a spatial filter must discard information to be useful. In other words, it cannot be invertible and maps from a high-dimensional (possibly infinite-dimensional) space $\mathcal{V}$ of turbulent fields to a lower-dimensional space $\mathcal{W}$ of LES fields. In particular, defining a filter $\bar{\cdot}$ also defines its null space $\mathcal{V}'$ and complement $\bar{\mathcal{V}}$. The non-invertibility of the filter guarantees that $\mathcal{V}'$ is nontrivial, providing an unambiguous definition of unresolvable scales as elements of $\mathcal{V}'$. The space $\mathcal{W}$ of LES fields is isomorphic to $\bar{\mathcal{V}}$. This holds for continuous fields as well as discretized fields. In what follows, we denote by $\mathcal{V}$ the set of all fine-grid resolvable fields, by $\mathcal{W}$ the set of all coarse-grid resolvable fields, and use $\bar{\cdot}$ to indicate the application of a coarse-graining operator.

Therefore, for any LES field w, there exists a distribution of turbulent fields u such that $\tilde{u} = w$. By imposing that any spatial statistic of the true solution is reproduced by the LES solution, it is shown that the ideal LES evolution must satisfy

$$\frac{\mathrm{d}w}{\mathrm{d}t} = \left\langle \widetilde{\frac{\mathrm{d}u}{\mathrm{d}t}} \middle| \tilde{u} = w \right\rangle, \tag{3}$$

where $\langle \cdot | \cdot \rangle$ denotes the conditional average. In this formulation, it is necessary to think of the turbulent field u as a single realization in an ensemble of turbulent fields. Thus, Eq. (3) describes the average filtered evolution of an ensemble of fields u, conditional to the filtered fields $\tilde{u}$ being equal to the LES field w. Eq. (3) generalizes the results of Adrian [54], who showed that modeling turbulence reduces to approximating conditional averages of statistics.

The ideal LES evolution is rewritten as

$$\frac{\partial w}{\partial t} = L_{\mathrm{LES}}(w) + m(w), \tag{4}$$

where L_{LES} is an approximation of of the right-hand side of the original evolution, for example obtained as a coarse numerical discretization. Throughout this paper, we will refer to the numerical method integrating L_{LES} as the low-fidelity solver. The ideal LES model $m(w)$ is consequently defined as

$$m(w) = \langle M_{\mathrm{LES}}(u) | \tilde{u} = w \rangle, \tag{5}$$

$$M_{\mathrm{LES}}(u) = \widetilde{\frac{\partial u}{\partial t}} - L_{\mathrm{LES}}(\tilde{u}). \tag{6}$$

The above result holds for general flow configurations, numerical methods, and filters. Nonetheless, it is evident that all these choices will influence the model in practical situations. The flow domain and physical parameters naturally determine the operator L. The chosen filter and numerical discretization, which in itself implicitly induces a filter, influence L_{LES} and thus also affect the ideal subgrid model. Specifically, M_{LES} differs from M in (2) since the former contains discretization errors in addition to the commutator error already present in M. However, equations ((5)–(6)) indicate how to construct the ideal model once knowledge of the full system is available. In particular, it suggests that measurements of the subgrid force $M(u)$ in (6) are suited to serve as input for data-driven models. These measurements require integrating the filtered solution in time using the low-fidelity solver, which ensures that effects of spatial and temporal discretization are also measured, and comparing this to the filtered evolution of the unfiltered solution. As a consequence, the closure model will depend on the adopted numerical method and resolution, which is a necessary feature for consistency of the closure model.

2.2. Ideal LES in discrete time

As an intermediate step towards stochastic data-driven LES models, we proceed by expressing ideal LES in discrete time. This amounts to defining the *ideal distribution* of which the mean is the ideal LES prediction. This formulation will aid the model description in Section 3.

As before, we treat a turbulent field as a single entity in an ensemble of turbulent fields and adopt a probabilistic notation as follows. We denote by π_U the distribution of unfiltered turbulent fields, where $\pi_U^n(u)$ describes the probability of a turbulent field u being the realization of the flow dynamics at time t^n. We define the high-fidelity forward flow map $\psi_{\Delta t}(u^n) := u^{n+1} = u^n + L_{\Delta t}(u^n)$, where the superscripts denote the time levels and $L_{\Delta t}(u^n)$ denotes high-fidelity numerical integration of u^n for Δt time units. Since

the $\psi_{\Delta t}$ is a deterministic map, we can describe the evolution of a distribution of fields as a transition kernel $\tau(v|u) = \delta\big(v - \psi_{\Delta t}(u)\big)$ acting on the distribution π_U^n as

$$\pi_U^{n+1}(v) = \int_{\mathcal{V}} \delta\big(v - \psi_{\Delta t}(u)\big)\pi_U^n(u)\,\mathrm{d}u, \tag{7}$$

where δ is the Dirac delta function.

Specifically, given an LES solution w^n, we define the ideal distribution at t^{n+1} as follows. We begin with the distribution π_V^n of unfiltered fields u that satisfy $\tilde{u} = w^n$. This distribution evolves according to

$$\pi_V^{n+1}(v) = \int_{\mathcal{A}} \delta\big(v - \psi_{\Delta t}(u)\big)\pi_V^n(u)\,\mathrm{d}u, \tag{8}$$

where $\mathcal{A} = \{u \in \mathcal{V} | \tilde{u} = w^n\}$. We use $\pi_V^{n+1}(v)$ to define the distribution $\pi_W^{n+1}(w)$ of filtered fields w that correspond to the unfiltered fields at t^{n+1},

$$\pi_W^{n+1}(w) = \int_{\mathcal{B}} \pi_V^{n+1}(v)\,\mathrm{d}v, \tag{9}$$

where $\mathcal{B} = \{v \in \mathcal{V} | \tilde{v} = w\}$. We refer to π_W^{n+1} as the ideal distribution at time t^{n+1}. The ideal LES prediction w^{n+1} is then given as the average field in this distribution,

$$w^{n+1} = \mathbb{E}_{\pi_W^{n+1}}[w] = \int_{\mathcal{W}} w\,\pi_W^{n+1}(w)\,\mathrm{d}w. \tag{10}$$

Equations ((8)–(10)) describe the following procedure. Given an LES prediction at time w^n, we would like to know the best approximation at time w^{n+1} as described by ideal LES. We thus consider the distribution $\pi_V^n(u)$ of unfiltered fields u at time t^n that all correspond to the same LES field w^n after applying the filter $\bar{\cdot}$. Each unfiltered field in this distribution is evolved in time according to the high-fidelity flow map in (8), leading to a distribution $\pi_V^{n+1}(v)$ of unfiltered fields at time t^{n+1}. To find the ideal LES prediction at this time, we have to compute the corresponding distribution $\pi_W^{n+1}(w)$ of filtered fields w as in (9) and subsequently compute its mean following (10).

2.3. Naive Monte Carlo approximation to ideal LES

Eq. (10) indicates that the ideal LES prediction w^{n+1} at time t^{n+1} is the mean of the ideal distribution π_W^{n+1}. Computing w^{n+1} thus requires the evaluation of the integral on the right-hand side of Eq. (10), which is generally intractable since it concerns an infinite number of fluid fields. Any feasible approximation of this integral will require using Monte Carlo methods, which in this case corresponds to computing the empirical mean of a distribution approximating π_W^{n+1}. A naive approximation to the ideal LES prediction w^{n+1} can therefore be obtained by initializing an ensemble $\{u_i^n\}$ with an empirical distribution that approximates π_V^n, and use this ensemble to subsequently approximate the integrals in Eqs. (8)–(10). This is summarized in Algorithm 1.

Algorithm 1 only serves to illustrate the concept of ideal LES and will not be used for actual prediction of flow fields in the current study. We refer to this algorithm as 'naive' as it quickly becomes computationally intractable. Since every realization in the Monte Carlo approximation needs to be integrated with the high-fidelity solver, the computational costs grow rapidly even when a modest number of ensemble members are used in the approximation. To remedy this, we explore in the next section the combination of ideal LES and data assimilation techniques to approximate the ideal distribution without having to integrate at high spatial resolution.

Algorithm 1 Naive approximation of ideal LES over one time step.

procedure NAIVE IDEAL LES(w^n, ψ, Δt, N, $\bar{\cdot}$)
  for $i = 1, \dots, N$ **do**
    $u_i^n \leftarrow \text{sample}(\pi_V^n)$ ▷ Draw i.i.d. samples from π_V^n, recall that $\tilde{u}_i^n = w^n$
    $v_i^{n+1} \leftarrow \psi_{\Delta t}(u_i^n)$ ▷ Integrate samples in time to approximate π_V^{n+1}
    $w_i^{n+1} \leftarrow \bar{v}_i^{n+1}$ ▷ Filter samples to approximate π_W^{n+1}
  end for
  $w^{n+1} \leftarrow \frac{1}{N}\sum_{i=1}^N w_i^{n+1}$ ▷ The empirical mean of π_W^{n+1} approximates the ideal LES prediction
  return w^{n+1}
end procedure

The modeling framework proposed in this paper aims to approximate ideal LES via ensemble predictions, rather than provide the exact deterministic evolution. The strict non-invertibility requirement of the filter can then be relaxed and one may instead also consider ill-conditioned invertible filters. In such cases, a perturbation in a filtered field is magnified when computing the unfiltered field. Thus, one can turn a distribution of minimally perturbed LES fields into fully resolved fields that are considerably different and subsequently consider the mean evolution of this distribution as an approximation to the ideal LES evolution.

3. Probabilistic closure modeling framework

The probabilistic description of ideal LES as presented in the previous section motivates using stochastic simulation strategies to obtain empirical approximations to distributions of fields. Finding probability distributions given observations of the underlying dynamical system has traditionally been the goal of data assimilation [35], which aims to combine model predictions with observational data in a manner that optimally balances the uncertainties present in both the predictions and the observations. This is commonly achieved by filtering, i.e., sequentially updating the probability distribution of the variable of interest in two steps referred to as the prediction and the analysis.

In this section, we formulate stochastic LES in the context of sequential data assimilation. In particular, we discuss applying a Bayesian correction to the stochastic LES prediction in order to approximate the ideal distribution. By considering flows in statistically stationary states, we relax this modeling criterion and focus on accurate prediction of flow statistics instead. Through this combination of ideal LES and data assimilation methods, we arrive at a general framework of closure modeling by assimilating statistics.

Throughout this section, we assume that high-fidelity data is available in the form of flow snapshots. From these snapshots, we can extract time series and empirical distributions of flow statistics (also referred to as quantities of interest or QoIs) and access this information in the model formulation.

3.1. Sequential data assimilation approximation to ideal LES

We first briefly recall the ensemble prediction and Bayesian correction steps in sequential data assimilation, which enables us to formulate the general closure model in Section 3.2.

Ensemble LES prediction. For simplicity, we consider a single LES solution w^n at time t^n and formulate the *prediction ensemble* or *forecast ensemble* starting from this single solution. As before, we can define the forward flow map as the (stochastic) map $\varphi_{\Delta t,i}(w^n) := w_i^{n+1} = w^n + L_{\mathrm{LES},\Delta t}(w^n, m_i)$, where $L_{\mathrm{LES},\Delta t}(w^n, m_i)$ denotes numerical integration of the LES solution including the i^{th} realization of the stochastic LES model. Thus, through the choice of numerical method and LES model, we implicitly define the transition kernel $\tau_{\mathrm{LES},\Delta t}(w_f|w)$. This kernel can be thought of as describing the probability that $\varphi_{\Delta t}(w) = w_f$ occurs, or in other words, that the flow configuration w_f is reached after integrating the prediction model Δt time units starting from configuration w. Using this notation, we define the forecast distribution π_f^{n+1} at time t^{n+1} as

$$\pi_f^{n+1}(w_f) = \int_{\mathcal{W}} \tau_{\mathrm{LES},\Delta t}(w_f|w)\delta(w - w^n)\,\mathrm{d}w. \tag{11}$$

If one defines the stochastic LES model $m(w)$ such that its average equals the conditional average in equations (5–6), then the mean of the stochastic ensemble will be the ideal LES prediction. However, in practice this will generally not be achievable. The conditional average in Eq. (5) can be estimated with sufficient measurements of $M(u)$ (Eq. (6)) using nonlinear mean square estimation [55]. Nonetheless, it becomes increasingly difficult to obtain sufficient samples of the state space when its dimension increases [36], hence it is infeasible to compute good approximations of the conditional average even at modest resolutions of the LES prediction. In practice, therefore, the forecast distribution π_f^{n+1} will generally differ from the ideal distribution and may even be markedly different if errors accumulate. We thus need to correct the forecast distribution based on available knowledge of the ideal distribution.

Bayesian correction to incorporate measurements. The produced forecast ensemble is not guaranteed to be an accurate approximation of the ideal distribution. Therefore, we incorporate known information of the ideal distribution π_W^{n+1} into the forecast distribution π_f^{n+1} in a Bayesian assimilation step, thereby producing an *analysis ensemble* or *posterior ensemble* with a distribution denoted by π_a^{n+1}. The assimilation can succinctly be written in terms of a transition kernel $b(w_a|w_f)$ acting on the forecast ensemble [36, Chapter 7],

$$\pi_a^{n+1}(w_a) = \int_{\mathcal{W}} b(w_a|w_f)\pi_f^{n+1}(w_f)\,\mathrm{d}w_f, \tag{12}$$

where the form of b depends on the adopted DA method.

As an example we consider the Ensemble Kalman filter (EnKF) with perturbed observations [38,39] as a corrector for the prediction ensemble. EnKF provides a practical algorithm since it defines the transition from a forecast distribution to an analysis distribution in terms of the corresponding ensembles. We denote by $w_{i,f}^{n+1}$ and $w_{i,a}^{n+1}$ the i^{th} ensemble members of the forecast ensemble and analysis ensemble, respectively, at time t^{n+1}. The correction then takes the form

$$w_{i,a}^{n+1} = w_{i,f}^{n+1} - K\left(H w_{i,f}^{n+1} + r_i^{n+1} - y_{\text{obs}}^{n+1}\right), \tag{13}$$

where $r_i^{n+1} \sim \mathcal{N}(0, R)$ and y_{obs}^{n+1} is an observation. Here, H is the measurement operator mapping an LES prediction $w_{i,f}^{n+1}$ to an observable, K is the Kalman gain matrix and R is the observational noise covariance matrix. For this method, the transition kernel b can be expressed as [36, Section 7.1]

$$b\left(w_a^{n+1}|w_f^{n+1}\right) = \mathrm{n}\left(w_a^{n+1}; w_f^{n+1} - K\left(H w_f^{n+1} - y_{\text{obs}}^{n+1}\right), KRK^T\right). \tag{14}$$

The term on the right-hand side describes the probability of sampling w_a^{n+1} from a normal distribution with mean $w_f^{n+1} - K(Hw_f^{n+1} - y_{\text{obs}}^{n+1})$ and covariance KRK^T.

3.2. Closure modeling by assimilating statistics

Using the assumption that the flow is statistically stationary and that high-fidelity flow snapshot data are available, we now derive a framework for ensemble methods that serve as self-contained data-driven stochastic models.

A flow statistic is defined by a function $G : \mathcal{W} \to \mathbb{R}$ producing a single number for a given flow configuration. Then, the expected value of the statistic in the ideal distribution

$$\mathbb{E}_{\pi_W}[G(w)] = \int_{\mathcal{W}} G(w)\pi_W(w)\,\mathrm{d}w \tag{15}$$

becomes time-independent since we assume statistical stationarity. Therefore, instead of focusing on approximating the ideal distribution π_V^{n+1} by the posterior distribution π_a^{n+1}, we can choose to relax this modeling goal such that it applies to selected flow statistics only. The underlying assumption is that selected flow statistics adequately describe the ideal distribution and hence that an ensemble of solutions with accurate flow statistics approximates the ideal distribution. That is, we aim to find a posterior distribution π_a^{n+1} of flow fields such that

$$\mathbb{E}_{\pi_a^{n+1}}[G(w)] = \mathbb{E}_{\pi_W}[G(w)] \tag{16}$$

for user-defined flow statistics G. This way, we shift our focus from predicting distributions of fields to predicting distributions of flow statistics. In doing so, we exploit the knowledge of the distribution of G as measured from high-fidelity flow snapshots. We denote this distribution by $G(w)\pi_W$.

To achieve good approximations of flow statistics, following (16), we consider sequential data assimilation applied to a predicted ensemble of statistics rather than an ensemble of flow fields. This gives rise to the following general method, also summarized in Algorithm 2. Starting from an initial field w^n, an LES prediction ensemble $\{w_{i,f}^{n+1}\}$ is computed from which flow statistics $\{G_{i,f}^{n+1}\}$ are extracted. A Bayesian method is subsequently employed to correct the predicted statistics. Namely, a DA method is applied to the predicted statistic ensemble $\{G_{i,f}^{n+1}\}$, where 'observations' $\{G_{i,\mathrm{obs}}^{n+1}\}$ are samples from the distribution $G(w)\pi_W$. The latter is estimated from the high-fidelity data. The resulting updated values $\{G_{i,a}^{n+1}\}$ of the statistic should be satisfied by the analysis ensemble of flow fields $\{w_{i,a}^{n+1}\}$. Thus, the final step of the algorithm is to find, or 'reconstruct', this ensemble. Specifically, each forecast ensemble member $w_{i,f}^{n+1}$ is transformed into an analysis ensemble member $w_{i,a}^{n+1}$ such that the latter satisfies the flow statistic $G_{i,a}^{n+1}$.

Algorithm 2 Framework for closure modeling by assimilating statistics over one time step.

procedure CLOSURE MODEL(w^n, φ, Δt, N, DA)
  for $i = 1, \dots, N$ **do**
    $w_{i,f}^{n+1} \leftarrow \varphi_{\Delta t,i}(w^n)$ ▷ Compute ensemble LES prediction
    $G_{i,f}^{n+1} \leftarrow G(w_{i,f}^{n+1})$ ▷ Compute predicted statistics
    $G_{i,\mathrm{obs}}^{n+1} \leftarrow \mathrm{sample}(G(w)\pi_W)$ ▷ Generate 'observed' statistics
  end for
  $\left\{G_{i,a}^{n+1}\right\} \leftarrow \mathrm{DA}\left(\left\{G_{i,f}^{n+1}\right\}, \left\{G_{i,\mathrm{obs}}^{n+1}\right\}\right)$ ▷ Bayesian correction of predicted statistics
  for $i = 1, \dots, N$ **do**
    $w_{i,a}^{n+1} \leftarrow \mathrm{reconstruct}(w_{i,f}^{n+1}, G_{i,a}^{n+1})$ ▷ Reconstruct flow fields such that these satisfy the updated statistics
  end for
  return $\left\{w_{i,a}^{n+1}\right\}$ ▷ Return the ensemble of flow fields with updated statistics
end procedure

Algorithm 2 provides a general framework for including knowledge of desired flow statistics into stochastic LES predictions. However, its generality comes at the cost of having many modeling choices, which we list below. These choices are

1. how to measure the sub-grid force $M(u)$ in Eq. (6). Ideal LES establishes what this measurement should encompass in continuous time. In discrete time, one has to decide over which time interval $M(u)$ is measured;
2. which stochastic LES model to use. Even though ideal LES provides a goal for what the LES model should achieve, the form of the stochastic forcing term may depend on the problem at hand;
3. which statistics to assimilate;
4. how the 'observations' are defined based on high-fidelity data. Time series data of statistics can be extracted from available high-fidelity flow snapshots. In turn, one can choose, e.g., to draw samples from empirically measured distributions, or to mimic measured temporal correlation;
5. which DA scheme to use;
6. how to reconstruct the flow fields when statistical values are known. A small number of statistics does not uniquely define a flow field, hence it is desirable to optimally find flow fields $w_{i,a}^{n+1}$ from $w_{i,f}^{n+1}$ under the constraints defined by the desired flow statistics.

The above description highlights that the reference data is used twice in this framework: first to measure the sub-grid force $M(u)$, which is indicative of the local integration error, and subsequently to determine the flow statistics, which describe the desired statistically steady state. Furthermore, the first two points are related to ensemble prediction with stochastic data-driven LES, the remainder are part of the Bayesian correction. In the next section, we illustrate how the presented modeling framework can be applied

to develop a data-driven stochastic closure for two-dimensional Rayleigh-Bénard convection, and elaborate on each of the required modeling choices.

4. Application to Rayleigh-Bénard convection

In this section, we provide an example of a closure model combining stochastic LES and assimilating statistics. We first introduce the governing equations and the test case that the closure model will be applied to. Subsequently, we highlight the model choices, provide implementation details and an estimate of the resulting computational complexity.

4.1. Governing equations and numerical methods

Two-dimensional Rayleigh-Bénard (RB) convection serves as the test bed for the proposed model. The governing equations are the incompressible Navier-Stokes equations coupled to an energy equation describing buoyancy effects under the Boussinesq approximation. The nondimensionalized equations are

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u} = \sqrt{\frac{Pr}{Ra}}\nabla^2\mathbf{u} - \nabla p + T\mathbf{e}_y, \tag{17}$$

$$\nabla\cdot\mathbf{u} = 0, \tag{18}$$

$$\frac{\partial T}{\partial t} + \mathbf{u}\cdot\nabla T = \frac{1}{\sqrt{PrRa}}\nabla^2 T, \tag{19}$$

where $\mathbf{u}$ is the velocity, p is the pressure, T is the temperature, and $\mathbf{e}_y$ is the unit vector in the vertical direction. The velocity consists of a horizontal velocity u_x and a vertical velocity u_y. Following the notation of Section 2, we regard a solution u as the pair $(\mathbf{u}, T)$.

The equations are solved in a two-dimensional rectangular box of width $L_x = 2$ and height $L_y = 1$. Periodic boundary conditions are imposed on the sides of the domain for all variables. The top and bottom boundaries are walls with no-slip boundary conditions for the velocity, and prescribed values of 1 at the bottom and 0 at the top for the dimensionless temperature. The dimensionless numbers are the Prandtl number $Pr = \nu/\kappa$ and the Rayleigh number $Ra = g\beta\Delta L_y^3/(\nu\kappa)$. These numbers describe the ratio of characteristic length scales of the velocity and the temperature and the ratio between buoyancy and viscous effects, respectively, and we have a Reynolds number $Re = \sqrt{Ra/Pr}$. Here ν is the kinematic viscosity, κ is the thermal diffusivity, g is the gravitational acceleration, β is the thermal expansion coefficient, Δ is the temperature difference between the walls of the domain. The test case is run at $Pr = 1$ and $Ra = 10^{10}$ to simulate in the turbulent convective regime.

The Nusselt number describes the heat flux from the bottom to the top of the domain and is one of the critical responses of the system to the imposed physical parameters [56]. We adopt the definition

$$Nu = 1 + \sqrt{PrRa}\langle vT\rangle_\Omega, \tag{20}$$

where $\langle\cdot\rangle_\Omega$ is the mean over the entire domain Ω. Definition (20) is particularly suited for computation on coarse grids since it does not involve any gradients, and will be used as a metric to assess coarse-grid simulations.

The equations ((17)–(19)) are discretized in space using a energy-conserving finite difference method [57]. The high-fidelity solver is parallelized following [58]. The velocity is defined on a staggered grid, the temperature is defined on the same grid as the vertical velocity, and the pressure is computed at the cell centers. The arrangement of the temperature and vertical velocity ensures no interpolation errors when computing the buoyancy term [59]. A hyperbolic tangent grid spacing is adopted in the vertical (wall-normal) direction, guaranteeing refinement near the walls to resolve the boundary layer. The grid is uniformly spaced in the horizontal direction.

The temporal discretization follows from a fractional-step third-order Runge-Kutta (RK3) for explicit terms and the Crank-Nicholson (CN) scheme for implicit terms. A time step from t^n to t^{n+1} is divided into three sub-stages denoted by the superscript $k, k = 0, 1, 2$, where $k = 0$ coincides with the situation at t^n. In each stage, a provisional velocity $\mathbf{u}^*$ is computed as

$$\frac{\mathbf{u}^* - \mathbf{u}^k}{\Delta t} = \left[\gamma_k H^k + \rho_k H^{k-1} - \alpha_k \mathcal{G}p^k + \alpha_k \mathcal{A}_y^k \frac{\mathbf{u}^* + \mathbf{u}^k}{2}\right]. \tag{21}$$

The Runge-Kutta coefficients γ, ρ, α are given by $\gamma = [8/15, 5/12, 3/4]$, $\rho = [0, -17/60, -5/12]$ and $\alpha = \gamma + \rho$ (see [58–60]). The discrete gradient is denoted by $\mathcal{G}$. The convective terms, horizontal diffusion terms and source terms (buoyancy) are collected in H^k and are treated explicitly. The vertical diffusion term $\mathcal{A}_y$ is treated implicitly to eliminate viscous stability restrictions arising from the non-uniform grid near the boundary [61]. A Poisson equation is then solved using $\mathbf{u}^*$ to impose the continuity constraint (18). Discretely, this is given by

$$\mathcal{L}\phi = \frac{1}{\alpha_k \Delta t}\left(D\mathbf{u}^*\right), \tag{22}$$

where $\mathcal{L}$ is the discrete Laplacian and D is the discrete divergence. The velocity and pressure are then updated to yield a divergence-free velocity field,

$$\mathbf{u}^{k+1} = \mathbf{u}^* - \alpha_k \Delta t(\mathcal{G}\phi), \tag{23}$$

$$p^{k+1} = p^k + \phi - \frac{\alpha_k \Delta t}{2Re}(\mathcal{L}\phi). \tag{24}$$

The QUICK interpolation scheme [62] is used to discretize the convective terms, whereas the diffusive terms are discretized with a standard second-order finite difference scheme in both spatial directions. Finite differences are used to define the discrete gradient $\mathcal{G}$, the discrete divergence $\mathcal{D}$ and the discrete Laplacian $\mathcal{L}$.

A resolution of 4096×2048 is adopted for the DNS, which has been shown to be sufficient to resolve the flow at the chosen Rayleigh number [63]. All subsequent coarse-grid simulations are carried out on a 64×32 grid. The flow cannot be fully resolved at this low resolution and the numerical method induces artificial dissipation [30,64]. Hence, the low-fidelity numerical solution will reach a different statistically steady state than the high-fidelity reference solution if no model is explicitly applied.

4.2. Model specification

As presented in Algorithm 2 in Section 3, the full statistical closure modeling framework consists of a stochastic ensemble forecast method and a subsequent Bayesian correction. A total of six modeling choices have to be made to obtain a workable algorithm, which we elaborate below. Model generalization is discussed in Section 6.

We assume that a sequence of high-fidelity snapshots $\{u^j\}$ is available. Furthermore, we repeatedly make use of the periodicity of the domain in the horizontal (wall-parallel) direction by computing the one-dimensional Fourier transform along horizontal cross-sections of the domain. The Fourier coefficients are indicated with a hat symbol $\hat{\cdot}$. For example, we use $\widehat{u_{x}}_{,k,l}$ to denote the k^{th} Fourier coefficient of the velocity u_x along the l^{th} horizontal cross-section. Given an entire solution u consisting of a velocity $\mathbf{u} = (u_x, u_y)$ and a temperature T, one obtains the Fourier coefficients $\widehat{u_{x}}_{,k,l}$, $\widehat{u_{y}}_{,k,l}$, $\widehat{T}_{k,l}$. We denote these coefficients collectively by $\hat{u}_{k,l}$ for readability. This leads to the following model description.

1. The subgrid force M in Eq. (6) is recorded in an *a posteriori* measurement. Specifically, $M(u^n)$ is measured by integrating the high-fidelity solver in time from a high-fidelity snapshot u^n and the low-fidelity solver from $\tilde{u}^n$. Both solutions are integrated for Δt time units, which equals a single coarse-grid time step and constitutes multiple fine-grid time steps. The solution from the high-fidelity solver is then filtered and compared to the solution of the low-fidelity solver. Thus, a set of measurements $\{M_{\text{LES},\Delta t}(u^j)\}$ is obtained from a collection of high-fidelity snapshots $\{u^j\}$. This procedure is summarized in Algorithm 3, where $\varphi_{\Delta t}(w^n)$ indicates numerical integration of the LES solution without an LES model. The number of measurements is varied throughout the performed numerical experiments and is specified in Section 5.

Algorithm 3 Single a posteriori measurement.

procedure A POSTERIORI MEASUREMENT(u^n, ψ, φ, Δt, $\tilde{\cdot}$)
  $w^n \leftarrow \tilde{u}^n$ ▷ Initialize low-fidelity solution as the filtered high-fidelity snapshot
  $u^{n+1} \leftarrow \psi_{\Delta t}(u^n)$ ▷ Integrate high-fidelity solution in time
  $w^{n+1} \leftarrow \varphi_{\Delta t}(w^n)$ ▷ Integrate low-fidelity solution in time
  $M_{\text{LES},\Delta t}(u^n) = \tilde{u}^{n+1} - w^{n+1}$
  return $M_{\text{LES},\Delta t}(u^n)$
end procedure

2. The LES model is implemented as a stochastic perturbation calibrated from the measurements $\{M_{\text{LES},\Delta t}(u^j)\}$. Namely, time series of the Fourier coefficients along horizontal cross-sections of the domain are first extracted for each prognostic variable. This yields $\{\widehat{M_{\text{LES},\Delta t}(u^j)}_{,k,l}\}$ and in particular we also obtain magnitudes $\{\text{abs}(\widehat{M_{\text{LES},\Delta t}(u^j)}_{,k,l})\}$. The sample means $\mu_{M,k,l}$ and variances $\sigma^2_{M,k,l}$ of the magnitudes are stored.

   The perturbation is then constructed as follows. For each Fourier coefficient and horizontal cross-section independently, we sample a magnitude $r_{M,k,l} \sim \mathcal{N}(\mu_{M,k,l}, \sigma^2_{M,k,l})$ and a phase $\vartheta_{M,k,l} \sim U[0, 2\pi]$. This fully defines the spectral coefficients and thus the corresponding physical fields. These fields are added as a perturbation after the full time step has been completed with the low-fidelity solver. An ensemble forecast $\{w^{n+1}_{i,f}\}$ is produced by computing a separate perturbation for each ensemble member. Such a perturbation can accurately compensate for coarsening errors even when a fraction of data is used [65] and is non-intrusive, thereby readily enabling ensemble forecasts.
3. We aim to reproduce the energy spectra and mean heat flux in all horizontal cross-sections of the domain. Given a solution u, the QoIs are $\text{abs}(\hat{u}_{k,l})$ for the energy spectra and $(\widehat{u_y T})_{0,l}$ for the mean heat flux, where the subscript 0 signifies the zeroth Fourier coefficient. For simplicity, we will denote a single QoI by G, which can refer to either the magnitude of a Fourier coefficient or the mean heat flux.
4. The 'observations' are based on the sequence of filtered high-fidelity snapshots $\{\tilde{u}^j\}$. This sequence yields a time series $\{G^j_{\text{ref}}\}$ for each QoI, of which the sample mean μ_{obs} and variance σ^2_{obs} are stored. At every assimilation step, an ensemble $\{G_{i,\text{obs}}\}$ of observations is generated by sampling $G_{i,\text{obs}} \sim \mathcal{N}(\mu_{\text{obs}}, \sigma^2_{\text{obs}})$ for each QoI separately.
5. A simplified ensemble Kalman filter (EnKF) [38] is employed to update the predicted statistics based on the observations. EnKF takes into account the nonlinearity of the governing equations, permits using a wide range of noise models [39], and is straightforward to implement.

   A diagonalization approach [41,66] is used that disregards covariances between the QoIs, so that the analysis acts on each QoI separately. This can be thought of as covariance 'localization'. Consequently, this avoids computing (inverses of) covariance matrices in the analysis step as is usually required in EnKF, which becomes a source of extensive computational costs if many

statistics are assimilated simultaneously. Instead, the analysis for each QoI reduces to a scalar equation independent of other QoIs. No ensemble inflation is used in the present study.

We denote the forecast, analysis ensemble, and observation ensembles of a single QoI by $\{G_{i,f}\}$, $\{G_{i,a}\}$, and $\{G_{i,\text{obs}}\}$, respectively. The diagonalized EnKF then updates each predicted QoI as

$$G_{i,a} = G_{i,f} + \left(\frac{\text{var}\left(\{G_{i,f}\}\right)}{\text{var}\left(\{G_{i,f}\}\right) + \text{var}\left(\{G_{i,\text{obs}}\}\right)} \right) \left(G_{i,\text{obs}} - G_{i,f}\right). \tag{25}$$

6. The predicted LES solution $w_{i,f}$ is updated to a solution $w_{i,a}$ satisfying the statistic $G_{i,a}$ from the previous step by appropriately adjusting the spectral coefficients of $w_{i,f}$. Namely, the magnitudes of the spectral coefficients of $w_{i,f}$ can be readily changed to $G_{i,a}$. The desired heat flux values are approximated by applying a gradient descent method on the phases of the Fourier coefficients of the temperature field, as developed in [30]. The solution $w_{i,a}$ is then fully described by the updated spectral coefficients.

The diagonalized EnKF in step 5 updates each QoI separately by disregarding covariances between the QoIs. Consequently, the error $|G_{i,a} - G_{i,\text{obs}}|$ can be estimated for each QoI independently, and is further detailed in Appendix A. In particular, for a sufficiently large ensemble size and sufficiently small time step, the ensemble mean of $\{G_{i,a}\}$ is shown to revert to μ_{obs}. Accurately predicting mean values of QoIs is a consistency requirement, since it is assumed that the specified QoIs adequately describe the ideal distribution.

4.3. Computational complexity

We now provide some estimates of the complexity reduction when employing the model on coarse grids. Solving the Poisson Eq. (22) is the largest source of computational costs in the spatial discretization since it involves solving a sparse $(N_x N_y) \times (N_x N_y)$ linear system. The high-fidelity solver employs a parallelization technique for finite-difference discretization of wall-bounded flows [59] which solves (22) in $\mathcal{O}(N_x N_y \log[N_y])$ operations. The low-fidelity solver uses a standard LU factorization for sparse linear systems for this purpose, requiring $\mathcal{O}((N_x N_y)^{3/2})$ computational steps [67].

The one-dimensional (inverse) fast Fourier transform (I(FFT)) along horizontal cross-sections of the domain is used extensively in the present adaptation of the model. These are computed in $\mathcal{O}(N_x \log[N_x])$ operations [68]. The random fields in the stochastic perturbations are computed by applying the IFFT at each horizontal profile for all prognostic variables. Applying the correction of the magnitudes of the Fourier coefficients requires computing the (I)FFT twice for these variables. The heat flux correction requires three computations of the (I)FFT in total and an additional $\mathcal{O}(N_x N_y)$ arithmetic operations per iteration. In total, we obtain the estimates of $\mathcal{O}(10^7)$ operations per time step for the high-fidelity solver, and $\mathcal{O}(10^5)$ for the low-fidelity solver including the model. This estimate is for a single realization of the model. Simulating an ensemble of size N increases the computational costs by approximately a factor N. The diagonalized EnKF (25) does not result in substantial additional computational costs.

The time step size becomes restrictive for high-Ra flow. A step size of $\mathcal{O}(\Delta y) \approx \mathcal{O}((RaNu)^{-1}) = \mathcal{O}(10^{-12})$ is suggested [59] to stably fully resolve the turbulent flow features. Here Δy denotes the smallest grid spacing in the non-uniform wall-normal grid.

The employed fine and coarse grids differ by a factor 64 in terms of grid cells per spatial direction. However, the difference between the smallest (non-uniform) grid spacings in the wall-normal direction yields a significant relaxation of the time step constraint. Following the estimate above, a time step restriction of $\mathcal{O}(10^{-4})$ is obtained on the adopted coarse grid. However, a time step size of $\mathcal{O}(10^{-2})$ was found to yield stable results, likely due to the presence of artificial dissipation, leading to a significant computational cost reduction.

5. Model performance assessment

This section presents simulation results on the coarse computational grid, comparing four models to the high-fidelity reference. We first introduce the employed coarse-grid models, the reference data used for model calibration, and the performance assessment metrics.

Coarse-grid models. Four different models are employed in the coarse-grid simulations and compared to the filtered DNS. These models are listed below and consist of a physics-based deterministic model and three ensemble methods.

- The no-model coarse numerical simulation serves as a deterministic physics-inspired closure model and will be referred to as the '*coarse, no model*' method. The artificial viscosity native to the coarse discretization was found to produce results similar to using a Smagorinsky eddy-viscosity model [69,70] at several Smagorinsky constants. The latter is not included in the results for brevity. In Section 5.4, we instead employ a conservative coarse-grid discretization to assess regularization properties of the model.
- The first ensemble method is a stochastic LES model without Bayesian correction, which we refer to as a '*stochastic LES model*'. The model uses only the first two steps described in Section 4.2. Hence, it relies only on the a posteriori measurements and does not require additional time series data of the QoIs. This model is a variant of the a posteriori statistical turbulence closure of [34], applied to the quantities of interest given in Section 4.2.
- The second ensemble method combines the coarse no-model simulation with a heuristic stochastic correction, and is referred to as a '*statistical nudge*'. Namely, an ad-hoc correction of the predicted QoIs is applied towards an observation following the work of [30]. The observation is defined as in the third step in Section 4.2, while the nudging strength is defined for each statistic independently through the measured correlation time [30]. This approach does not use the a posteriori measurements but instead only requires time series data of the QoIs.

- The third ensemble method combines stochastic LES with a Bayesian correction, using all steps described in Section 4.2, and will be denoted as '*stochastic LES model, assimilated*'.

The ensembles will consist of 10 members in Sections 5.1, 5.2, and 5.4. An ensemble with 50 members is considered in Section 5.3 to investigate the dependence of the prediction quality on the ensemble size.

Reference data. The reference data is a sequence of filtered snapshots of a DNS performed on a 4096×2048 computational grid, from which two data sets are extracted.

- The first data set is comprised of '*plenty data*'. A total of 1000 solution snapshots separated by 0.05 time units are used to compute the a posteriori measurements and compute the statistics used in the observations. The corresponding results are presented in Sections 5.1, 5.3, and 5.4.
- The second data set consists of '*few data*' and is used to verify the robustness of the model in the sparse data regime. Here, a total of 20 solution snapshots separated by 0.5 time units are used in the model calibration. The results using this data set are presented in Section 5.2.

The initial conditions of all coarse-grid simulations are filtered DNS snapshots outside the data sets described above.

Performance metrics. Several metrics are used to assess the model performance. For short lead times, a predicted solution $(\mathbf{u}, T)$ can be compared to the reference solution $(\mathbf{u}_{\text{ref}}, T_{\text{ref}})$ using the pattern correlation. We adopt the definition

$$\frac{\langle (\mathbf{u}, T), (\mathbf{u}_{\text{ref}}, T_{\text{ref}}) \rangle}{\sqrt{\langle (\mathbf{u}, T), (\mathbf{u}, T) \rangle \langle (\mathbf{u}_{\text{ref}}, T_{\text{ref}}), (\mathbf{u}_{\text{ref}}, T_{\text{ref}}) \rangle}}, \tag{26}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard L^2 inner product over the computational domain.

Long-time simulation results are evaluated via time-averaged energy spectra and root-mean-square deviations (r.m.s.), as well as rolling means of the kinetic energy (KE) and the Nusselt number. The r.m.s. is computed as a function of the wall-normal distance. For a given y-value, we compute the r.m.s. of a field f as

$$\text{r.m.s.}(f, y, t) = \left[\frac{1}{|A|} \int_A \left(f(x, y, t) - \langle f(x, y, t) \rangle_A \right)^2 \mathrm{d}A \right]^{1/2}, \tag{27}$$

where $\langle \cdot \rangle_A$ is the mean over a cross-section with length $|A|$. The KE is defined as

$$\text{KE} = \int_\Omega \frac{1}{2} (\mathbf{u} \cdot \mathbf{u}) \, \mathrm{d}\Omega. \tag{28}$$

The Nusselt number is computed following Eq. (20).

5.1. Model performance with plenty data

We first assess the model performance when plenty data is used to estimate the model parameters.

The pattern correlations (26) at short lead times are shown in Fig. 1 for three distinct initial conditions. The no-model simulation deteriorates over time due to loss of details in the numerical solution after initializing the flow from a filtered DNS snapshot. A low correlation does not imply that the flow statistics are incorrect since it measures the likeness of the global solution to the reference. Rather, it indicates the rate at which the flow deviates from the reference. It allows us to distinguish whether a solution is too rapidly steered towards a configuration with the desired statistical flow features. The correlation of the statistical nudging ensemble with the reference decreases rapidly, from which we surmise that the ad-hoc nudging is too strong and does not lead to a physical flow evolution. Applying the stochastic perturbation in the forecasting step largely mitigates this. This outcome reflects a trade-off between pointwise agreement with the reference solution and the accuracy of statistical properties. The correlation coefficient

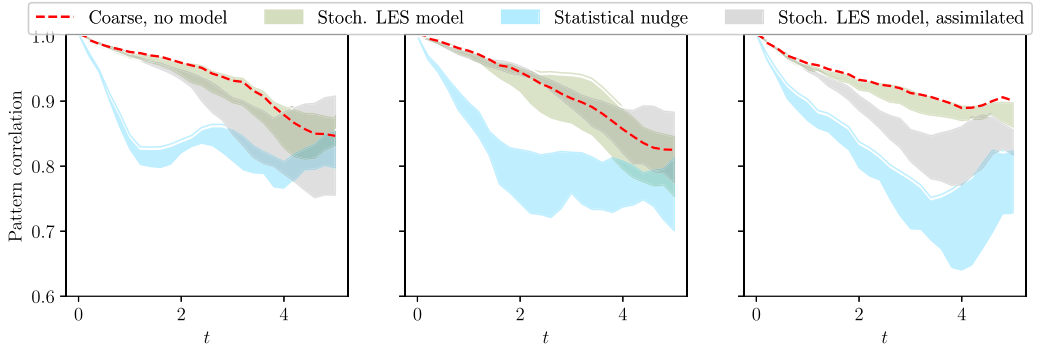


Fig. 1. Pattern correlation between the prediction and the reference solution, using plenty data to calibrate the model. Three different initial conditions are considered. Each ensemble consists of 10 members; each band is colored between the maximal and minimal measured values.

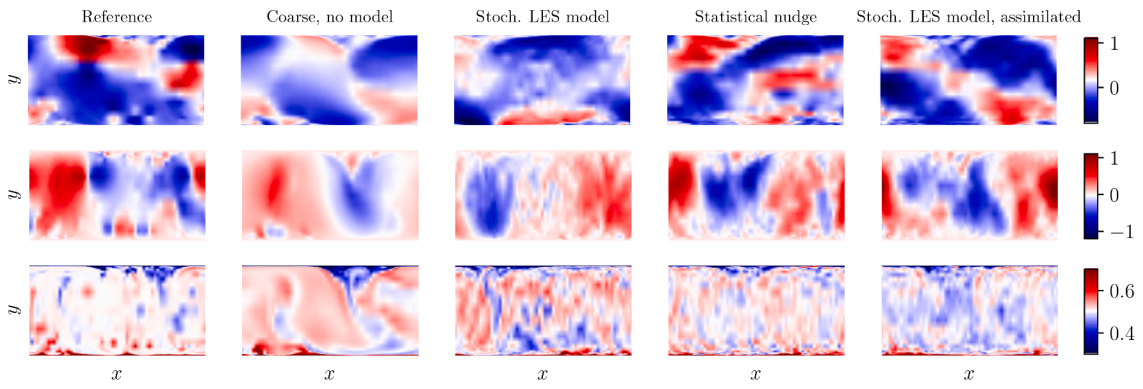


Fig. 2. Instantaneous snapshots after simulating for 100 time units initialized from a filtered DNS snapshot. Shown are the horizontal velocity (top row), vertical velocity (middle row) and temperature (bottom row). A snapshot from a single ensemble member is shown for the ensemble methods, using plenty data to calibrate the models.

primarily measures the pointwise agreement with a single reference, while statistical agreement does not necessarily imply strong pointwise correlation. Namely, a given set of statistics generally does not uniquely define the solution, hence a model can produce statistically accurate flow fields that are not well correlated with the reference. However, the difference between the 'statistical nudge' ensemble and the 'stochastic LES model, assimilated' ensemble suggests that a reduced nudging strength might be favorable when the initial condition is known. For example, this can be achieved by incorporating temporal effects into the observations, thereby including information of the initial conditions in the observation. Alternatively, a dynamically adaptive nudging procedure [44] can be applied such that a nudge is only applied if it increases the likelihood of the solution being in the ideal distribution.

A qualitative model comparison is given in Fig. 2 through instantaneous flow snapshots after reaching a statistically steady state in a long-time simulation. We observe that the no-model simulation suffers from artificial dissipation, evident from the smoothed fields and the reduced velocity magnitudes. The level of detail in the velocity fields is reconstructed well with the ensemble methods. The temperature fields display small-scale details, but these are fragmented instead of forming coherent patterns. Simultaneously, the velocity fields of the nudged ensembles feature pronounced flow details and maintain a qualitative agreement with the reference.

The average energy spectra near the center of the domain are depicted in the top row of Fig. 3, where the average is taken over all snapshots and all ensemble members. We also display the average spectra of the high-fidelity measurements that comprise the data set from which the model parameters are extracted and refer to this as the historical data. The no-model simulation provides predictions that consistently contain too little energy. Only applying the stochastic perturbation does not adequately alleviate this. Applying a correction, either ad-hoc or by assimilating statistics, yields a significant improvement particularly at the largest resolvable scales. This suggests that incorporating high-fidelity data into the model benefits the prediction of flow statistics in long-time simulations. We observe that the resulting energy spectra approximate the historical data accurately, which itself slightly deviates from the reference spectra over the simulated interval.

The time-averaged r.m.s. deviations show considerable improvement when applying the correction procedure, as shown in the bottom row of Fig. 3. Only applying the stochastic perturbation yields no improvement over using no model in the long-time simulations, which further highlights the added value of the statistical correction. Assimilating statistics leads to a pronounced improvement of the velocity r.m.s. deviation particularly near the center of the domain. Here, the grid size is largest and the discretization effects significant, which induce a large measured sub-grid force and thus large stochastic perturbations. In turn, these large perturbations lead to a larger gain factor in the correction (25) and a close adherence to the reference statistics.

The rolling averages of the total kinetic energy and the Nusselt number over time are respectively given in Fig. 4. The artificial dissipation in the coarsened discretization causes the no-model result to deviate from the filtered DNS and reach a different statistically steady state with a strongly reduced energy content. Only applying the stochastic perturbation does not alleviate this, whereas including high-fidelity statistical data substantially improves the total energy content. An improvement of the predicted Nusselt number is observed for all ensemble methods. Notably, this includes the ensemble in which the predictions are perturbed and no knowledge of the heat flux is included. However, judging from the energy spectra in Fig. 3, this result is obtained without correctly predicting the energy distributions in the velocity and temperature fields.

5.2. *Model performance with few data*

We now turn our attention to the model performance when using few data to estimate the model parameters. Only 20 snapshots are used to measure the the sub-grid scale forcing and the reference statistics. As such, the means and variances used in the model are poorly estimated and the correlation times of the quantities of interest become difficult to estimate due the sparsity of the data. We therefore expect the quality of the ad-hoc statistical nudging method to decrease.

The pattern correlations for short lead times are given in Fig. 5. No significant change in prediction quality is observed, compared to the predictions based on plenty data in Fig. 1. This suggests that the prediction of the instantaneous solution at short lead times is robust under changes in the available data and instead relies on the accuracy of the initial condition.

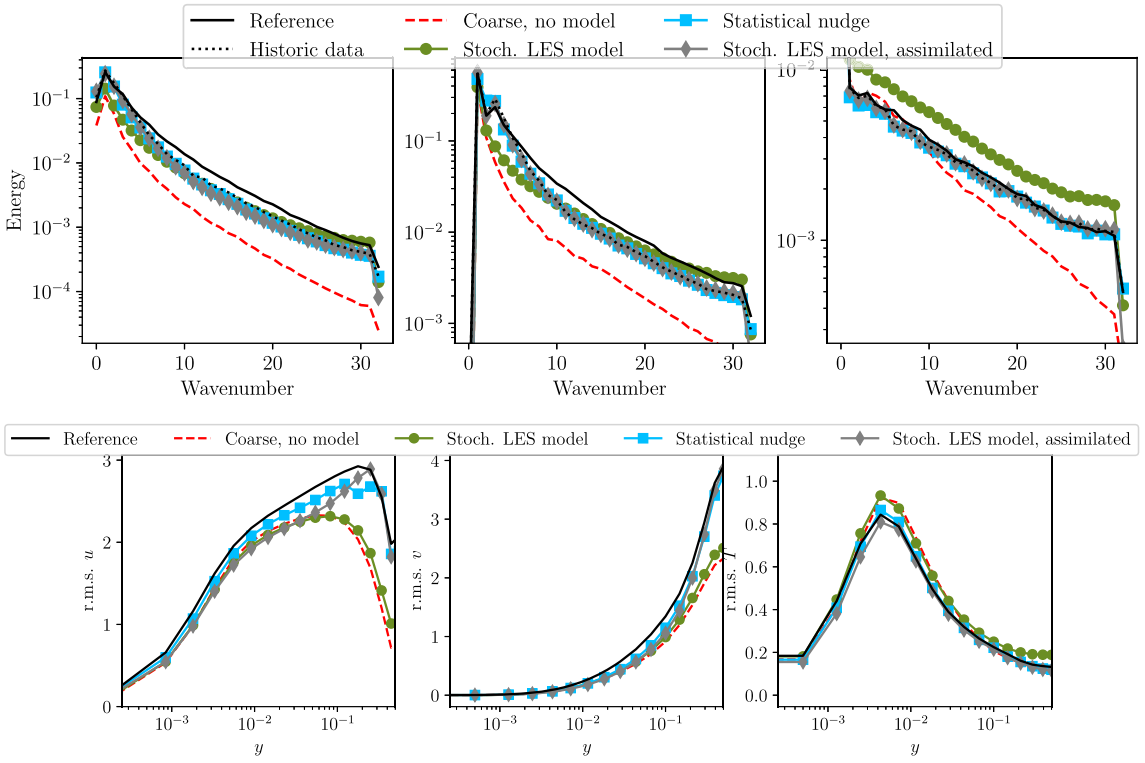


Fig. 3. Top: Time-averaged energy spectra measured along a horizontal cross-section of the domain for the horizontal velocity (left), vertical velocity (middle) and temperature (right). The cross-sections are take in the core of the domain at $y = 5.5 \times 10^{-1}$ for the horizontal velocity and $y = 5.0 \times 10^{-1}$ for the vertical velocity and the temperature. Bottom: Average root mean square deviation (r.m.s.) of the horizontal velocity (left), vertical velocity (middle) and temperature (right), measured along horizontal cross-sections of the domain and shown as functions of the wall-normal distance. The models are calibrated using plenty data.

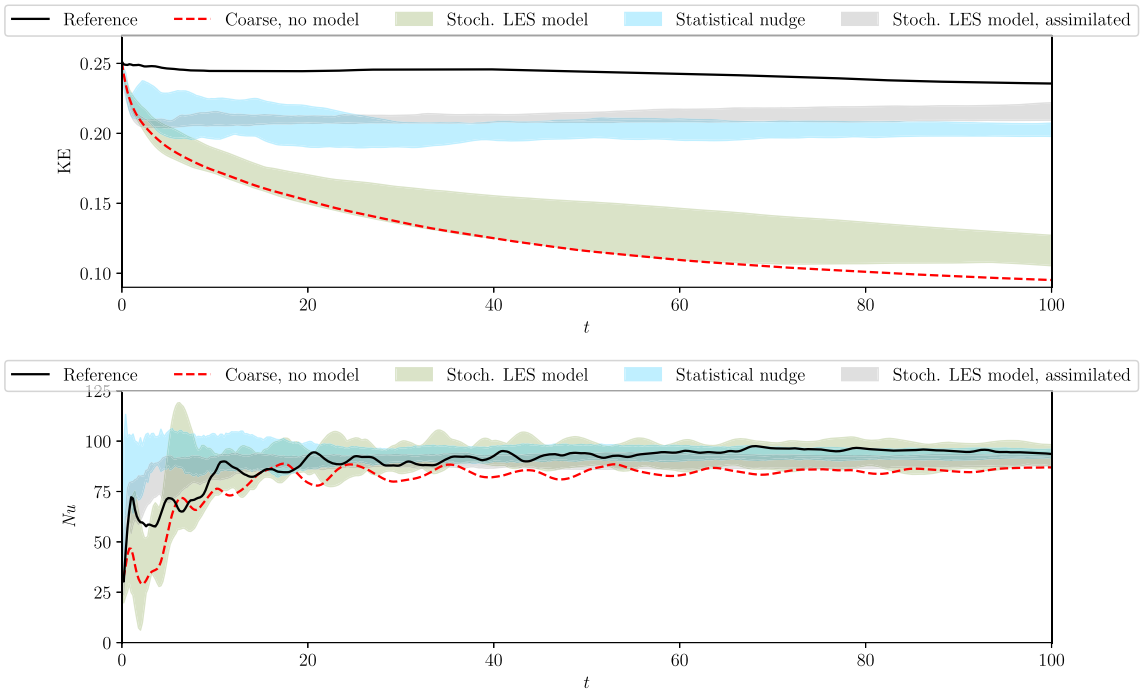


Fig. 4. Rolling means of the kinetic energy (KE, top panel) and the Nusselt number (Nu, bottom panel) over time, where the models are calibrated using plenty data. Each ensemble consists of 10 members; each band is colored between the maximal and minimal measured values.

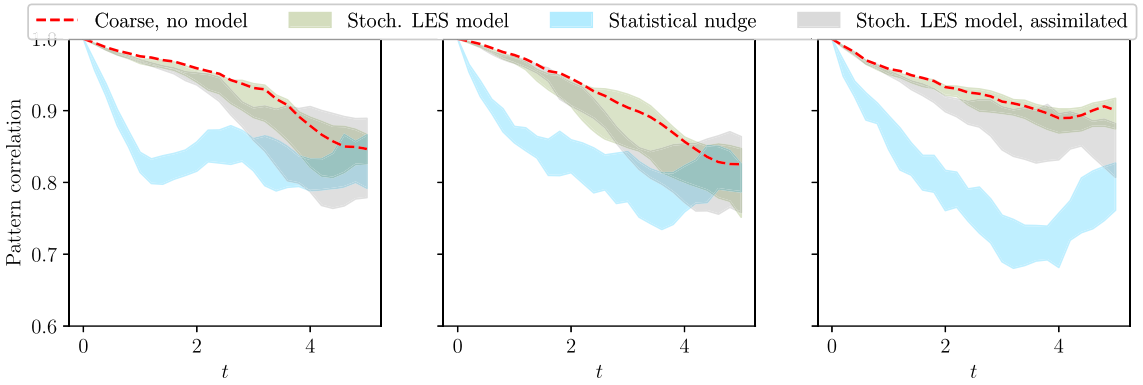


Fig. 5. Pattern correlation between the prediction and the reference solution, using few data to calibrate the model. Three different initial conditions are considered. Each ensemble consists of 10 members; each band is colored between the maximal and minimal measured values.

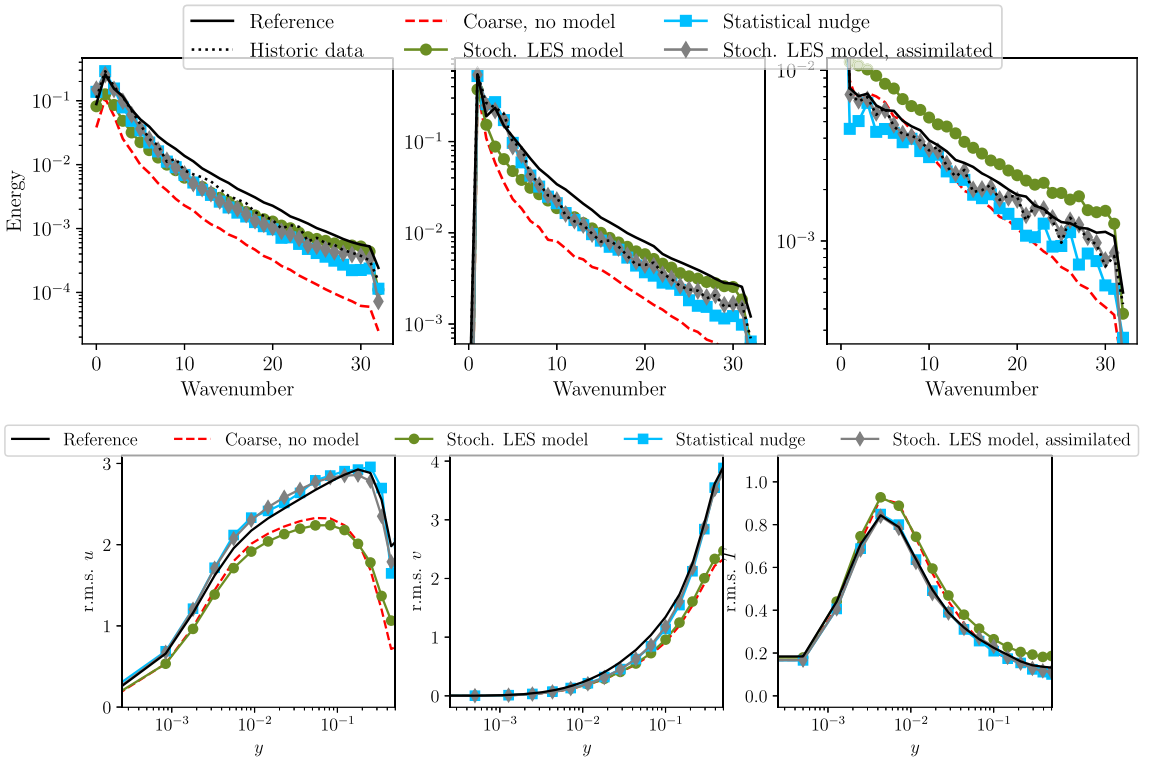


Fig. 6. Top: Time-averaged energy spectra measured along a horizontal cross-section of the domain for the horizontal velocity (left), vertical velocity (middle) and temperature (right). The cross-sections are take in the core of the domain at $y = 5.5 \times 10^{-1}$ for the horizontal velocity and $y = 5.0 \times 10^{-1}$ for the vertical velocity and the temperature. Bottom: Average root mean square deviation (r.m.s.) of the horizontal velocity (left), vertical velocity (middle) and temperature (right), measured along horizontal cross-sections of the domain and shown as functions of the wall-normal distance. The models are calibrated using few data.

The energy spectra in the core of the domain and the average r.m.s. values for the prognostic variables are depicted in Fig. 6. Good agreement is observed at the largest scales of motion despite the small amount of data, and the methods that employ a statistical correction adequately reproduce the energy in these scales. The effects of using limited data become apparent in the smaller scales of motion, particularly visible in the spectrum of the temperature. The average measured energy in these scales is not converged and hence deviate from the reference. The ensemble with assimilated statistics closely follows the historical data, for which an improvement over the no-model result is still evident. The average r.m.s. values do not deteriorate using the small data set and show good agreement with the reference. The rolling averages of the kinetic energy and the Nusselt number, shown in Fig. 7, display the same qualitative behavior is observed as when using plenty data. Overall, no distinct loss of predictive quality is found using few data to calibrate the model when compared to using plenty data.

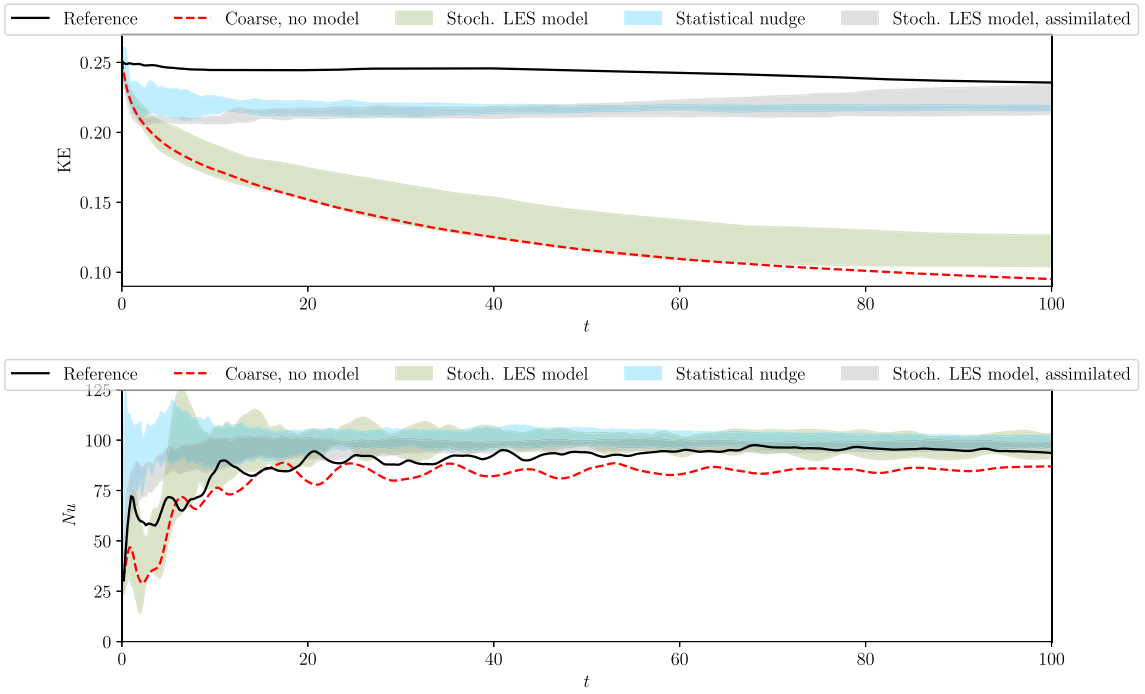


Fig. 7. Rolling means of the kinetic energy (KE, top panel) and the Nusselt number (Nu, bottom panel) over time, where the models are calibrated using few data. Each ensemble consists of 10 members; each band is colored between the maximal and minimal measured values.

5.3. Dependence on ensemble size

The EnKF results in the optimal linear estimator in the limit of large ensemble size [71,72]. Carrying out an ensemble simulation with many ensemble members might quickly become prohibitively expensive, even when using a computationally cheap low-fidelity solver. In the EnKF, the evolution of the individual ensemble members is coupled through the analysis step which uses information of the entire ensemble, and a change in system dynamics may thus be observed when changing the number of ensemble members. We therefore repeat the short-time numerical simulations presented in Fig. 1 in Section 5.1 using 50 ensemble members instead of 10 to establish that a modest ensemble size does not adversely affect the model performance. The obtained pattern correlations in Fig. 8 show no qualitative change with respect to the earlier presented results. This suggests that the currently adopted approach already provides robust forecasts at small ensemble size.

5.4. Regularization properties

We now turn to the regularizing features of the proposed model. As highlighted in the previous subsections, the QUICK interpolation scheme introduces significant artificial dissipation in coarse-grid simulations, which complicates the assessment of whether the

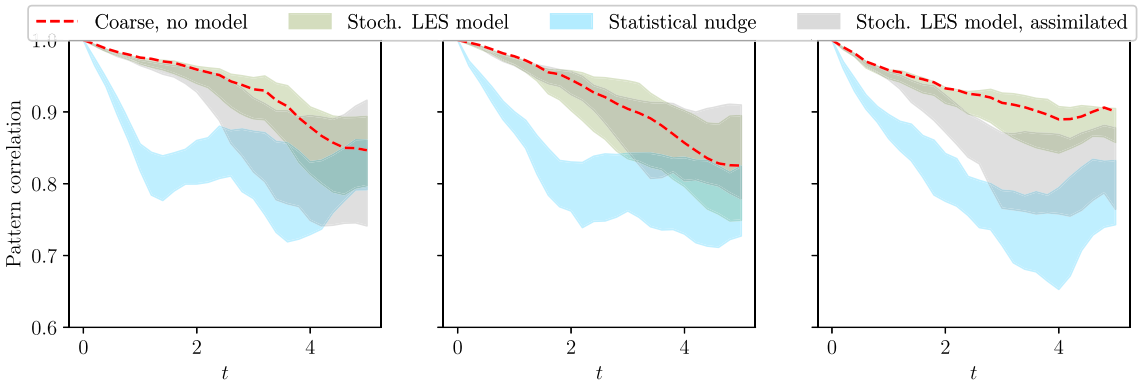


Fig. 8. Pattern correlation between the prediction and the reference solution, using plenty data to calibrate the model. Each ensemble consists of 50 members; each band is colored between the maximal and minimal measured values.

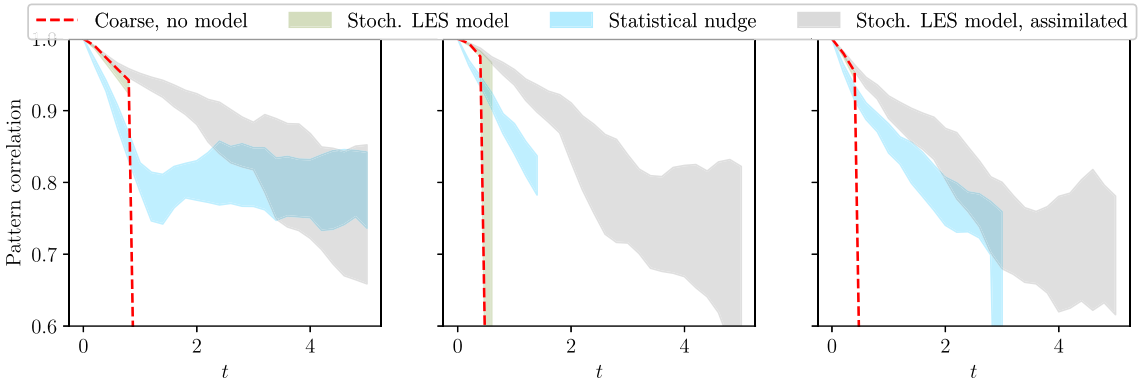


Fig. 9. Pattern correlation between the prediction and the reference solution, using plenty data to calibrate the model. A conservative central difference scheme is employed for the velocity advection on the coarse grid, and three different initial conditions are considered. Each ensemble consists of 10 members; each band is colored between the maximal and minimal measured values. Only the stochastic LES model with assimilated statistics remains stable.

proposed model itself contributes meaningful dissipative effects. To further investigate the regularization capabilities of the model, we repeat the numerical experiment from Section 5.1, this time employing a conservative central scheme for velocity advection [57]. In coarse-grid simulations, this scheme develops high-frequency instabilities unless either the time step is significantly reduced or a strongly dissipative sub-grid scale model is used.

The short lead-time pattern correlations are shown in Fig. 9. It is observed that numerically stable simulations are only consistently achieved when employing the stochastic LES model while assimilating statistics. The other methods break down due to numerical instability. Nonetheless, a lower pattern correlation is observed when compared to the results in 5.1.

These observations are also reflected in the long-time results depicted in Figs. 10 and 11. A striking agreement with the historical data is found for the energy spectra, suggesting that the model retains its capability to maintain the desired energy distribution across

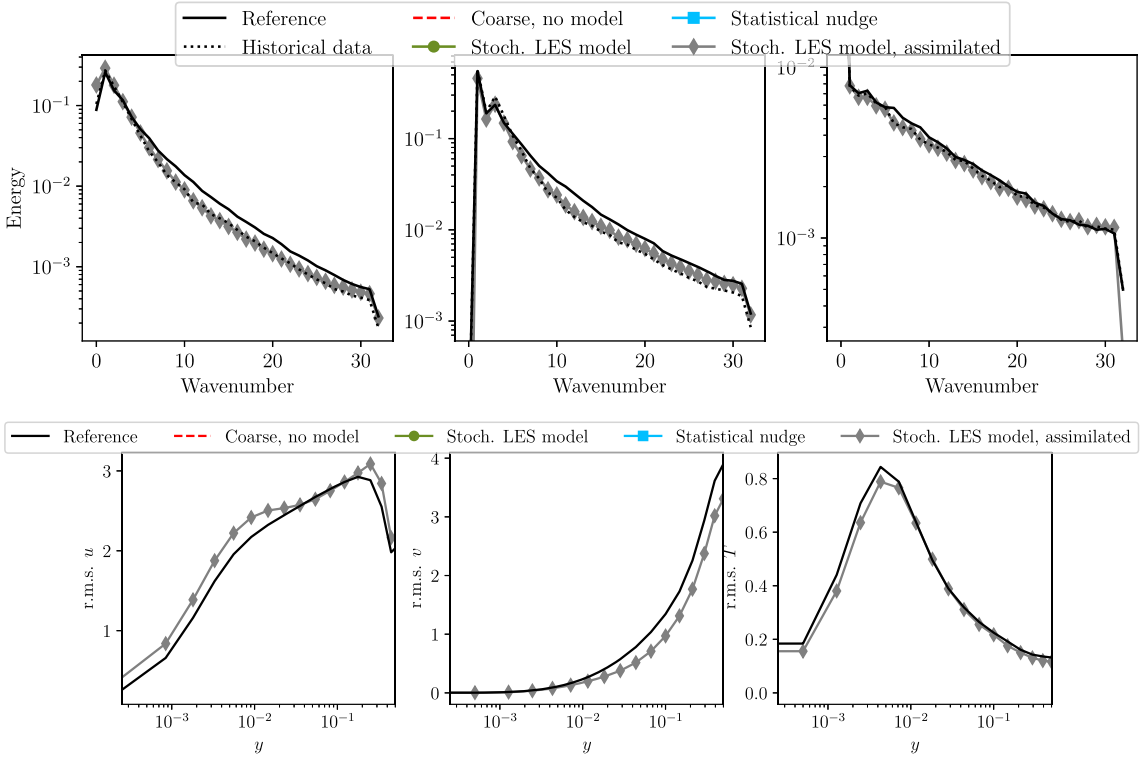


Fig. 10. Top: Time-averaged energy spectra measured along a horizontal cross-section of the domain for the horizontal velocity (left), vertical velocity (middle) and temperature (right). The cross-sections are take in the core of the domain at $y = 5.5 \times 10^{-1}$ for the horizontal velocity and $y = 5.0 \times 10^{-1}$ for the vertical velocity and the temperature. Bottom: Average root mean square deviation (r.m.s.) of the horizontal velocity (left), vertical velocity (middle) and temperature (right), measured along horizontal cross-sections of the domain and shown as functions of the wall-normal distance. The models are calibrated using few data, and a conservative central difference scheme is employed for the velocity advection on the coarse grid. Only the stochastic LES model with assimilated statistics remains stable and is shown.

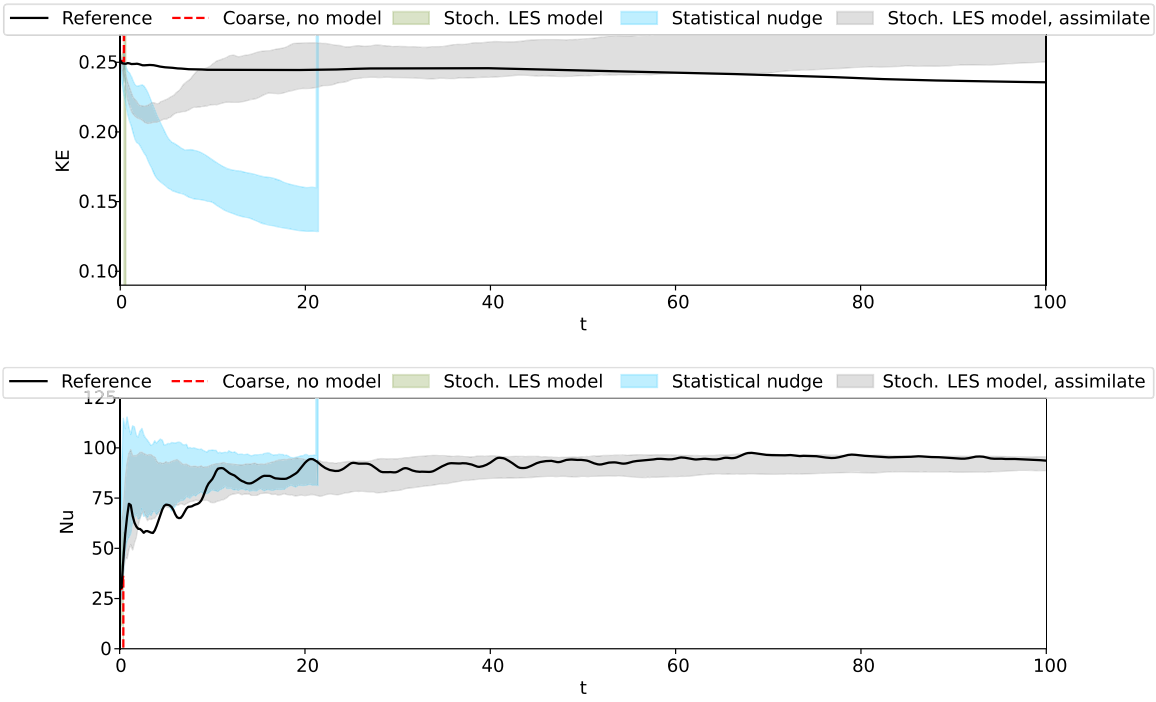


Fig. 11. Rolling means of the kinetic energy (KE, top panel) and the Nusselt number (Nu, bottom panel) over time, where the models are calibrated using few data and a conservative central difference scheme is employed for the velocity advection on the coarse grid. Each ensemble consists of 10 members; each band is colored between the maximal and minimal measured values. Only the stochastic LES model with assimilated statistics remains stable.

the resolvable scales of motion. This indicates that the model regularizes the solution and adapts its dissipative properties to the discretization effects, although it does not act as a dissipative term in the traditional sense. The r.m.s. profiles provide a reasonable approximation of the reference solution, albeit with reduced accuracy compared to those obtained using the QUICK scheme. The total kinetic energy is slightly overestimated while an accurate estimate of the Nusselt number is obtained. Despite the stabilizing features of the model illustrated here, a slight deterioration compared to the results in Section 5.1 is observed. This suggests that the proposed model might benefit from the simultaneous use of dissipative discretizations or LES models.

6. Discussion on model generalization

The model outlined in Section 4 serves as an example of how the general framework of closure modeling by assimilating statistics, described in Section 3, can be used in practice. Given the numerous modeling choices involved, we now discuss several potential directions for generalizing the model.

The sub-grid force measurements are not limited to one coarse-grid time step and may alternatively be used to model the instantaneous sub-grid force (6) to facilitate adaptive time-integration methods. Additionally, measuring the sub-grid force through more elaborate *a posteriori* methods (see [10] and references therein) can further aid the quality of the ensemble LES prediction.

The stochastic LES model is necessary to compute a forecast ensemble, and a wide range of modeling choices is available for this purpose. The present choice of de-correlated forcing of spectral coefficients intends to maintain simplicity and computational tractability, although incorporating spatial correlation in the forcing enables more refined forecast models. The empirical sub-grid force covariance is shown and discussed in Appendix B, and indicates that the covariance between high-frequency modes is non-negligible. A tractable way to include these covariances is via localization [36] in spectral space. It is worth noting that the stochastic LES model need not be designed to exactly mirror the chosen QoIs. For example, POD-based forcing can account for spatial correlations in a low-rank setting. Other stochastic models to approximate the ideal distribution include multiplicative noise to incorporate state-dependent forcing, structure-preserving stochastic perturbations to respect conserved quantities, or machine-learned parametrizations of unresolved dynamics.

Similar to the stochastic LES model, the 'observations' consist of de-correlated measurements of Fourier coefficients. This approach corresponds to the diagonalized EnKF method [41,66], which neglects interactions between spectral modes in favor of computational efficiency. However, as shown in Appendix B, the low-frequency components of the reference data exhibit non-negligible covariance with neighboring frequencies, both within horizontal cross-sections and across adjacent cross-sections. This is attributed to the coupling of Fourier modes through nonlinear advection, as well as the presence of coherent flow structures that span across multiple

horizontal layers. This finding suggests that the current method can be made more accurate by incorporating these covariance values in the observations.

A possible extension to three-dimensional wall-bounded setups, such as turbulent channel flow or Rayleigh-Bénard convection, can be addressed analogously to the presented two-dimensional case. For simple domains with structured grids, this could be achieved by reconstructing two-dimensional spectra in wall-parallel planes. Reconstructing POD spectra provides an alternative [31], which is suited to three-dimensional flows and complex geometries. At the same time, the number of tracked statistics can increase substantially in three-dimensional settings and a reduction of the number of QoIs might become necessary for the algorithm to remain computationally feasible. Within the presented modeling framework, the choice of QoIs remains flexible and may depend on the problem at hand. For instance, one may consider only low-wavenumber Fourier modes to capture dominant large-scale structures or restrict attention to horizontal planes near the walls to better resolve boundary layer dynamics. An alternative strategy involves the use of spatially integrated QoIs, such as the total kinetic energy or global heat flux, from which reduced computational models subject to physical constraints can be derived [33,34].

7. Conclusions

In this paper, we have proposed a method for deriving probabilistic data-driven turbulence closure models suitable for coarsened steady-state turbulence. Based on ideal large-eddy simulation, a combination of stochastic large-eddy similation and data assimilation methods is suggested, requiring both *a posteriori* collected data of the employed numerical solver and statistical data of the high-fidelity solution. Thus, the method exploits knowledge of the local integration error and the desired flow statistics.

An example of how the modeling framework can be used was provided using a non-intrusive implementation applied to two-dimensional Rayleigh-Bénard convection at Rayleigh number $Ra = 10^{10}$. Stochastic perturbations based on sub-grid scale data were used in conjunction with a simplified ensemble Kalman filter to steer coarse numerical predictions towards desired statistics known from a precursor simulation. Here, we focused on horizontal energy spectra and average heat flux, which were found to be accurately reproduced with the proposed model. The model showed robust results for a modest ensemble size. No considerable deterioration of the predictions was observed using as few as 20 high-fidelity snapshots to determine the model parameters. The model was found to accurately maintain the energy distributions observed in reference data, even when using an unstable discretization.

The presented modeling framework is general and can be applied to different fluid dynamical models. Nonetheless, many modeling choices have to be made in the actual implementation. An overview of the necessary modeling choices and suggestions for generalization were provided, including strategies to apply the general framework to three-dimensional flows. Exploring different modeling choices is a challenge for future work.

Here, the closure modeling framework based on the assimilation of statistics is formulated in discrete time. This formulation arises naturally in discretized systems and aligns with how data assimilation algorithms are typically presented. However, a continuous-time formulation can be advantageous for mathematical analysis (see, e.g., [72]) and aligns with the original ideal LES formulation [8]. The proposed modeling approach could likely be extended to a continuous-time setting, potentially offering new insights.

CRediT authorship contribution statement

Sagy R. Ephrati: Writing - review & editing, Writing - original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization.

Data availability

The data and adopted implementations that support the findings of this study are publicly available in Zenodo at http://doi.org/10.5281/zenodo.13353273.

Declaration of interests

The authors report no conflict of interest.

Acknowledgements

This work was supported by the Swedish Research Council (VR) through grant no. 2022-03453.

Appendix A. Error estimates for the adopted model

Below, we provide error estimates for the model described in Section 4.2.

Notation and definitions. The following estimate concerns a single quantity of interest G. Since each quantity of interest is updated independently of the other quantities, the analysis can be applied to each quantity of interest separately.

Let G^n and G^{n+1} be the quantity of interest at time instances t^n and $t^{n+1} = t^n + \Delta t$, respectively. The evolution of G is denoted by $\dot{G} = L_G(G)$. We consider an ensemble of N members, where each member is denoted by a subscript $i, i = 1, \dots, N$. We assume the deterministic operator L_G is Lipschitz continuous with Lipschitz constant C_G. Thus, we can use

$$G_i^n - \Delta t C_G \leq G_i^{n+1} \leq G_i^n + \Delta t C_G, \tag{A.1}$$

for each ensemble member.

The stochastic perturbation is given by $\Delta t \mu_m + \sigma_M r_{i,M}^n$, where $r_{i,M}^n \sim \mathcal{N}(0,1)$ and μ_m is the mean measured deviation after an integration step, specified as a model parameter. The observation at time instance n is denoted by $G_{i,\text{obs}}^n = \mu_{\text{obs}} + \sigma_{\text{obs}} r_{i,\text{obs}}^n$, where $r_{i,\text{obs}}^n \sim \mathcal{N}(0,1)$. The values of $r_{i,M}^n$ and $r_{i,\text{obs}}^n$ are i.i.d. for each i and n.

The predicted value of ensemble member i is given by $G_{i,f}$,

$$G_{i,f} = \int_{t^n}^{t^{n+1}} L_G(G_i^n)\,\mathrm{d}t + \Delta t \mu_m + \sigma_M r_{i,M}^n \tag{A.2}$$

We denote the gain by K and observe that this value is the same for each ensemble member. The ensemble and the observations have nonzero variance, from which it follows that $0 < K < 1$. We expand

$$\begin{aligned} G_i^{n+1} &= G_{i,f} + K(G_{i,\text{obs}}^{n+1} - G_{i,f}) \\ &= (1-K)\left(\int_{t^n}^{t^{n+1}} L(G_i^n)\,\mathrm{d}t + \Delta t \mu_m + \sigma_M r_{i,M}^n \right) + K G_{i,\text{obs}}^{n+1}, \end{aligned} \tag{A.3}$$

which is used to bound the errors of individual ensemble members and corresponding observations, as well as the ensemble mean and the mean observation.

Error between individual ensemble members and measurements. We define the errors as the distance between the value of the quantity of interest at a time instance and the corresponding observation, that is, $E_i^{n+1} = G_i^{n+1} - d_i^{n+1}$ and $E_i^n = G_i^n - d_i^n$. We find

$$E_i^{n+1} = G_i^{n+1} - d_i^{n+1} = (1-K)\left(\int_{t^n}^{t^{n+1}} L(G_i^n)\,\mathrm{d}t + \Delta t \mu_m + \sigma_M r_{i,M}^n - G_{i,\text{obs}}^{n+1} \right). \tag{A.4}$$

Note that $G_{i,\text{obs}}^{n+1} - G_{i,\text{obs}}^n = \sigma_{\text{obs}}\left(r_{i,\text{obs}}^n - r_{i,\text{obs}}^n \right)$ and therefore, using (A.1), we obtain

$$\begin{aligned} \left| E_i^{n+1} \right| &< \left| G_i^n + \Delta t C_G + \Delta t \mu_m + \sigma_M r_{i,M}^n - G_{i,\text{obs}}^{n+1} \right| \\ &\leq \left| G_i^n + \Delta t C_G + \Delta t \mu_m + \sigma_M r_{i,M}^n - G_{i,\text{obs}}^n - \sigma_{\text{obs}}\left(r_{i,\text{obs}}^{n+1} - r_{i,\text{obs}}^n \right) \right| \\ &= \left| E_i^n + \Delta t C_G + \Delta t \mu_m + \sigma_M r_{i,M}^n - \sigma_{\text{obs}}\left(r_{i,\text{obs}}^{n+1} - r_{i,\text{obs}}^n \right) \right| \\ &\leq \left| E_i^n \right| + \Delta t \left| C_G + \mu_m \right| + \sigma_M \left| r_{i,M}^n \right| + \sigma_{\text{obs}} \left| r_{i,\text{obs}}^{n+1} - r_{i,\text{obs}}^n \right|. \end{aligned} \tag{A.5}$$

Error between ensemble mean and mean observation. Recall the ensemble mean of a quantity f with ensemble members f_i, $i = 1, \dots, N$,

$$\bar{f} = \frac{1}{N} \sum_{i=1}^{N} f_i. \tag{A.6}$$

We now study the development of distance between the ensemble mean of the quantity of interest and the measured mean value, i.e., $\bar{E}^{n+1} = |\bar{G}^{n+1} - \mu_{\text{obs}}|$. We find that

$$\begin{aligned} E^{n+1} &= |\bar{G}^{n+1} - \mu_{\text{obs}}| \\ &= \left| \frac{1}{N} \sum_{i=1}^{N} (1-K) \left[\int_{t^n}^{t^{n+1}} L(G_i^n)\,\mathrm{d}t + \Delta t \mu_m + \sigma_M r_{i,M}^n - \mu_{\text{obs}} \right] + K \sigma_{\text{obs}} r_{i,\text{obs}}^{n+1} \right| \\ &\leq (1-K) \left| \frac{1}{N} \sum_{i=1}^{N} G_i^n - \mu_{\text{obs}} \right| + (1-K) \left| \frac{1}{N} \sum_{i=1}^{N} \Delta t C_G + \Delta t \mu_m + \sigma_M r_{i,M}^n \right| + K \sigma_{\text{obs}} \left| \frac{1}{N} \sum_{i=1}^{N} r_{i,\text{obs}}^{n+1} \right| \\ &< \left| \bar{G}^n - \mu_{\text{obs}} \right| + \Delta t \left(C_G + \mu_m \right) + \sigma_M \left| \overline{r_{i,M}^n} \right| + \sigma_{\text{obs}} \left| \overline{r_{i,\text{obs}}^{n+1}} \right| \\ &= E^n + \Delta t \left(C_G + \mu_m \right) + \sigma_M \left| \overline{r_{i,M}^n} \right| + \sigma_{\text{obs}} \left| \overline{r_{i,\text{obs}}^{n+1}} \right| \end{aligned} \tag{A.7}$$

Since $r_{i,M}^n$ and $r_{i,\text{obs}}^n$ are i.i.d. with mean zero for each i, n, we observe that $\overline{r_{i,M}^{n+1}}$ and $\overline{r_{i,\text{obs}}^{n+1}}$ approach zero as the ensemble size N grows. Thus, in the limit of infinite ensemble size, we obtain

$$E^{n+1} < E^n + \Delta t (C_G + \mu_m). \tag{A.8}$$

Eq. (A.8) shows that the ensemble mean of G reverts to μ_{obs} when the ensemble size is sufficiently large and the time step is sufficiently small.

Appendix B. Measured covariances between quantities of interest

The measured covariances between the magnitudes of the Fourier coefficients are shown in Fig. B.1, displaying the logarithm of the absolute value of the covariance to aid the identification of coherent patterns. Each horizontal cross-section of the domain consists of 33 Fourier coefficients. These are denoted by indices $0, 1, 2, \ldots, 1056$. Indices 0, 1, and 2 respectively correspond to the first, second, and third coefficients in the bottom-most horizontal layer, and index 1056 corresponds to the highest frequency in the top-most layer. The separate treatment of horizontal cross-sections causes a distinct squared pattern in the covariance matrices. Additionally, the vertical velocity and temperature have fixed boundary values, leading to a zero covariance with all other variables. These are visualized as white bands.

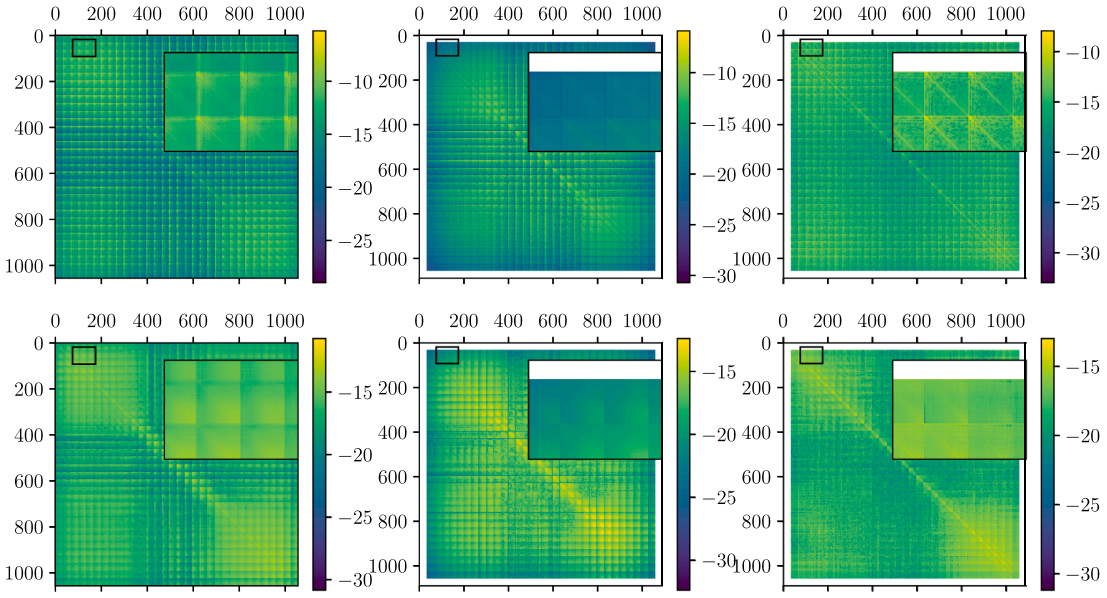


Fig. B.1. Logarithm of the absolute value of the covariance between the magnitudes of the Fourier coefficients. A total of 33 Fourier coefficients are computed along horizontal cross-sections of the domain, giving rise to the squared pattern. The figure displays the covariance for the horizontal velocity (left), vertical velocity (middle), and temperature (right), computed for the reference data (top row) and the measured sub-grid force (bottom row).

A diagonally dominant pattern is observed, both in the entire covariance matrix and the minor squares within. The covariance of the reference data furthermore indicates that the low-frequency components have non-negligible covariance with variables both within a horizontal cross-section and in neighboring layers. This is attributed to the coupling of Fourier modes via nonlinear advection, as well as the presence of coherent flow structures that span across multiple horizontal layers. For the sub-grid force, a similar pattern is observed for the high-frequency components.

References

[1] S.B. Pope, Turbulent flows. Meas. Sci. Technol. 12 (11) (2001) 2020–2021.
[2] W. Snyder, C. Mou, H. Liu, O. San, R. Devita, T. Iliescu, Reduced order model closures: a brief tutorial, in: Recent Advances in Mechanics and Fluid-Structure Interaction with Applications: The Bong Jae Chung Memorial Volume, Springer, 2022, pp. 167–193.
[3] B.J.Geurts, Direct and Large-Eddy Simulation, vol. 1, Walter de Gruyter GmbH & Co KG, 2022.
[4] P. Sagaut, Large Eddy Simulation for Incompressible Flows: An Introduction, Springer Science & Business Media, 2005.
[5] B.J. Geurts, D.D. Holm, Alpha-modeling strategy for LES of turbulent mixing, Turbulent Flow Computation, pp. 237–278, 2002.
[6] U. Piomelli, A. Rouhi, B.J. Geurts, A grid-independent length scale for large-eddy simulations, J. Fluid Mech. 766 (2015) 499–527.
[7] A. Rouhi, U. Piomelli, B.J. Geurts, Dynamic subfilter-scale stress model for large-eddy simulations, Phys. Rev. Fluids 1 (4) (2016) 44401.
[8] A. Jacob, R.D. Langford, Moser, Optimal LES formulations for isotropic turbulence, J. Fluid Mech. 398 (1999) 321–346.
[9] A. Beck, D. Flad, C.-D. Munz, Deep neural networks for data-driven LES closure models, J. Comput. Phys. 398 (2019) 108910.
[10] B. Sanderse, P. Stinis, R. Maulik, S.E. Ahmed, Scientific machine learning for closure models in multiscale problems: a review, arXiv preprint arXiv:2403.02913, 2024.
[11] S.D. Agdestein, Agdestein, B. Sanderse, Discretize first, filter next: learning divergence-consistent closure models for large-eddy simulation, arXiv preprint arXiv:2403.18088, 2024.
[12] K. Duraisamy, Perspectives on machine learning-augmented Reynolds-averaged and large eddy simulation models of turbulence, Phys. Rev. Fluids 6 (5) (2021) 50504.
[13] H. Frezat, J.L. Sommer, R. Fablet, G. Balarac, R. Lguensat, A posteriori learning for quasi-geostrophic turbulence parametrization, J. Adv. Model. Earth Syst. 14, (11) (MS003124) (2022).
[14] R. Buizza, M. Milleer, T.N. Palmer, Stochastic representation of model uncertainties in the ECMWF ensemble prediction system, Q. J. R. Meteorol. Soc. 125 (560) (1999) 2887–2908.
[15] K. Hasselmann, Stochastic climate models Part I. Theory. Tellus, 28, (6), 473–485, 1976.
[16] Palmer, Stochastic weather and climate models, Nat. Rev. Phys. 1 (7) (2019) 463–471.

[17] N.E. Lorenz, Predictability: a problem partly solved, Proc. Seminar on Predictability vol. 1 (1996).
[18] H. Arnold, Moroz, Palmer, Stochastic parametrizations and model uncertainty in the Lorenz'96 system, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 371 (1991) 20110479.
[19] D. Crommelin, E. Vanden-Eijnden, Subgrid-scale parameterization with conditional Markov chains, J. Atmos. Sci. 65 (8) (2008) 2661–2675.
[20] A. Boral, Z.Y. Wan, L.Z.-N. nez, J. Lottes, Q. Wang, Y.-F. Chen, J. Anderson, F. Sha, Neural ideal large eddy simulation: modeling turbulence with neural stochastic differential equations, Adv. Neural Inf. Process. Syst. 36 (2023) 69270–69283.
[21] D.D. Holm, Variational principles for stochastic fluid dynamics, Proc. R. Soc. A Math. Phys. Eng. Sci. 471 (2015) 20140963.
[22] E. Mémin, Fluid flow dynamics under location uncertainty, Geophys. Astrophys. Fluid Dyn. 108 (2) (2014) 119–146.
[23] C. Cotter, D. Crisan, D.D. Holm, W. Pan, I. Shevchenko, Numerically modeling stochastic Lie transport in fluid dynamics, Multiscale Model. Simul. 17 (1) (2019) 192–232.
[24] P.S.R. Ephrati, E. Cifani, B.J. Luesink, Geurts, Data-driven stochastic Lie transport modeling of the 2D Euler equations, J. Adv. Model. Earth Syst. 15 (MS003268) (2022).
[25] V. Resseguier, W. Pan, B. Fox-Kemper, Data-driven versus self-similar parameterizations for stochastic advection by Lie transport and location uncertainty, Nonlinear Process. Geophys. 27 (2020) 209–234.
[26] C. Cotter, D. Crisan, D.D. Holm, W. Pan, I. Shevchenko, A particle filter for stochastic advection by Lie transport: a case study for the damped and forced incompressible two-dimensional Euler equation, SIAM/ASA J. Uncertainty Quantif. 8 (4) (2020) 1446–1492.
[27] J.S. Frederiksen, V. Kitsios, T.J. O'Kane, Statistical dynamics and subgrid modelling of turbulence: from isotropic to inhomogeneous, arXiv preprint arXiv:2407.10085, 2024.
[28] J.S. Frederiksen, V. Kitsios, T.J. O'Kane, M.J. Zidikheri, Stochastic subgrid modelling for geophysical and three-dimensional turbulence, in: Nonlinear and Stochastic Climate Dynamics, Cambridge University Press, 2017, pp. 241–275.
[29] G. Berkooz, P. Holmes, J.L. Lumley, The proper orthogonal decomposition in the analysis of turbulent flows, Annu. Rev. Fluid Mech. 25 (1) (1993) 539–575.
[30] P.S.R. Ephrati, B.J. Cifani, Geurts, Data-driven spectral turbulence modelling for Rayleigh-Bénard convection, J. Fluid Mech. 975 (2023).
[31] S. Ephrati, Cifani, Geurts, Stochastic data-driven POD-based modeling for high-fidelity coarsening of two-dimensional Rayleigh-Bénard turbulence, in: ER-COFTAC Workshop Direct and Large Eddy Simulation, Springer, 2023, pp. 209–214.
[32] S. Jorgen, S.M. Frederiksen, Kepert, Dynamical subgrid-scale parameterizations from direct numerical simulations, J. Atmos. Sci. 63 (11) (2006) 3006–3019.
[33] W. Edeling, D. Crommelin, Reducing data-driven dynamical subgrid scale models by physical constraints, Comput. Fluids 201 (2020) 104470.
[34] R. Hoekstra, D. Crommelin, W. Edeling, Reduced data-driven turbulence closure for capturing long-term statistics, Comput. Fluids 285 (2024) 106469.
[35] K. Law, A. Stuart, K. Zygalakis, Data Assimilation, vol. 214, 52 Springer, Cham, Switzerland, 2015.
[36] S. Reich, C. Cotter, Probabilistic Forecasting and Bayesian Data Assimilation, Cambridge University Press, 2015.
[37] A.B.V. Rosić, J. Kučerová, O. Sykora, A. Pajonk, H.G. Litvinenko, Matthies, Parameter identification in a probabilistic setting, Eng. Struct. 50 (2013) 179–196.
[38] G. Evensen, Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res. Oceans, 99, C5, 1994, pp. 10143–10162.
[39] G. Evensen, The ensemble Kalman filter: theoretical formulation and practical implementation, Ocean Dyn. 53 (2003) 343–367.
[40] A. Bain, D. Crisan, Fundamentals of Stochastic Filtering, vol. 3, Springer, 2009.
[41] A.J. Majda, J. Harlim, Filtering Complex Turbulent Systems, Cambridge University Press, 2012.
[42] E. Bach, T. Colonius, I. Scherl, A. Stuart, Filtering dynamical systems using observations of statistics, Chaos 34, (3) 2024.
[43] R. Villiers, V. Mons, D. Sipp, E. Lamballais, M. Meldi, Enhancing unsteady reynolds-averaged Navier-Stokes modelling from sparse data through sequential data assimilation and machine learning. Flow Turbul. Combust., 1–39, 2025.
[44] Ömer Deniz, Akyildiz, J. Míguez, Nudging the particle filter, Stat. Comput. 30 (2020) 305–330.
[45] A. Azouani, E. Olson, E.S. Titi, Continuous data assimilation using general interpolant observables, J. Nonlinear Sci. 24 (2014) 277–304.
[46] E. Carlson, A. Larios, E.S. Titi, Super-exponential convergence rate of a nonlinear continuous data assimilation algorithm: the 2D Navier-Stokes equation paradigm, J. Nonlinear Sci. 34 (2) (2024) 37.
[47] M. Gesho, E. Olson, E.S. Titi, A computational study of a data assimilation algorithm for the two-dimensional Navier-Stokes equations, Commun. Comput. Phys. 19 (4) (2016) 1094–1110.
[48] M. Altaf, T. Titi, O.M. Gebrael, L. Knio, Zhao, I. Mccabe, Hoteit, Downscaling the 2D Bénard convection equations using continuous data assimilation, Comput. Geosci. 21 (2017) 393–410.
[49] M.A. E.R. Hammoud, O.L. Maître, S. Edriss, I. Titi, O. Hoteit, Knio, Continuous and discrete data assimilation with noisy observations for the Rayleigh-Bénard convection: a computational study, Comput. Geosci. 27 (1) (2023) 63–79.
[50] D. Blömker, K. Law, A.M. Stuart, K.C. Zygalakis, Accuracy and stability of the continuous-time 3DVAR filter for the Navier-Stokes equation, Nonlinearity 26 (8) (2013) 2193.
[51] P. Courtier, W. Andersson, D. Heckley, M. Vasiljevic, Hamrud, F. Hollingsworth, M. Rabier, Fisher, Pailleux, The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: formulation, Q. J. R. Meteorol. Soc. 124 (550) (1998) 1783–1807.
[52] S.R. Ephrati, P. Cifani, M. Viviani, B.J. Geurts, Data-driven stochastic spectral modeling for coarsening of the two-dimensional Euler equations on the sphere, Phys. Fluids 35 (9)2023.
[53] A.S.R. Ephrati, E. Franken, P. Luesink, B.J. Cifani, Geurts, Continuous data assimilation closure for modeling statistically steady turbulence in large-eddy simulation, Phys. Rev. Fluids 10 (1) (2025) 13801.
[54] R.J. Adrian, On the role of conditional averages in turbulence theory, Symposia on Turbulence in Liquids, 34 1975.
[55] A. Papoulis, S.U. Pillai, Probability, Random Variables, and Stochastic Processes, McGraw-Hill Europe, New York, NY, USA, 2002.
[56] G. Ahlers, S. Grossmann, D. Lohse, Heat transfer and large scale dynamics in turbulent Rayleigh-Bénard convection, Rev. Mod. Phys. 81 (2) (2009) 503–537.
[57] A. Vreman, The projection method for the incompressible Navier-Stokes equations: the pressure near a no-slip wall, J. Comput. Phys. 263 (2014) 353–374.
[58] P. Cifani, Kuerten, Geurts, Highly scalable DNS solver for turbulent bubble-laden channel flow, Comput. Fluids 172 (2018) 67–83.
[59] P. Erwin, V. Der, R. Poel, J. Ostilla-Mónico, R. Donners, Verzicco, A pencil distributed finite difference code for strongly turbulent wall-bounded flows, Comput. Fluids 116 (2015) 10–16.
[60] M.M. Rai, P. Moin, Direct simulations of turbulent flow using finite-difference schemes, J. Comput. Phys. 96 (1) (1991) 15–53.
[61] J. Kim, P. Moin, Application of a fractional-step method to incompressible Navier-Stokes equations, J. Comput. Phys. 59 (2) (1985) 308–323.
[62] B.P. Leonard, A stable and accurate convective modelling procedure based on quadratic upstream interpolation, Comput. Methods Appl. Mech. Eng. 19 (1) (1979) 59–98.
[63] X. Zhu, V. Mathai, P. Richard, R. Stevens, D. Verzicco, Lohse, Transition to the ultimate regime in two-dimensional Rayleigh-Bénard convection, Phys. Rev. Lett. 120 (14) (2018) 144502.
[64] J. Bernard, F. Geurts, V.D. Bos, Numerically induced high-pass dynamics in large-eddy simulation, Phys. Fluids 17 (12) (2005).
[65] E.S.R. Ephrati, G. Luesink, P. Wimmer, B.J. Cifani, Geurts, Computational modeling for high-fidelity coarsening of shallow water equations based on subgrid data, Multiscale Model. Simul. 20 (4) (2022) 1468–1489.
[66] J. Harlim, Majda, Filtering nonlinear dynamical systems with linear stochastic models, Nonlinearity 21 (6) (2008) 1281.
[67] A. George, E. Ng, On the complexity of sparse QR and LU factorization of finite-element matrices, SIAM J. Sci. Stat. Comput. 9 (5) (1988) 849–861.
[68] G.H. Golub, C.F Van. [Loan], Matrix Computations, JHU Press, 2013.
[69] S.-H. Peng, L. Davidson, Comparison of subgrid-scale models in LES for turbulent convection flow with heat transfer, Turbulent Heat Transf. 2 (1998) 5–24.
[70] J. Smagorinsky, General circulation experiments with the primitive equations: I. The basic experiment, Mon. Weather Rev. 91 (3) (1963) 99–164.
[71] J.H. Kody, H. Law, R. Tembine, Tempone, Deterministic mean-field ensemble Kalman filtering, SIAM J. Sci. Comput. 38 (3) (2016) 1251–A1279.
[72] C. Schillings, A.M. Stuart, Analysis of the ensemble Kalman filter for inverse problems, SIAM J. Numer. Anal. 55 (3) (2017) 1264–1290.