Attractor Learning in Synchronized Chaotic Systems in the Presence of Unresolved Scales[a]

W. Wiegerinck[1, 2, b] and F.M. Selten[3, c]

[1] *SNN Adaptive Intelligence, Nijmegen, The Netherlands*

[2] *Department of Biophysics, Donders Institute for Brain, Cognition and behavior, Radboud University Nijmegen, The Netherlands*

[3] *Royal Netherlands Meteorological Institute, De Bilt, The Netherlands*

(Dated: 12 Oct. 2017)

Recently, supermodels consisting of an ensemble of interacting models, synchronizing on a common solution, have been proposed as an alternative to the common non-interactive multi-model ensembles in order to improve climate predictions. The connection terms in the interacting ensemble are to be optimized based on data. The supermodel approach has been successfully demonstrated in a number of simulation experiments with an assumed ground truth and a set of good, but imperfect models. The supermodels were optimized with respect to their short-term prediction error. Nevertheless, they produced long-term climatological behavior that was close to the long-term behavior of the assumed ground truth, even in cases where the long-term behavior of the imperfect models was very different. In these supermodel experiments, however, a perfect model class scenario was assumed, in which the ground truth and imperfect models belong to the same model class and only differ in parameter setting. In this paper we consider the imperfect model class scenario, in which the ground truth model class is more complex than the model class of imperfect models due to unresolved scales. We perform two supermodel experiments in two toy problems. The first one consists of a chaotically driven Lorenz 63 oscillator ground truth and two Lorenz 63 oscillators with constant forcings as imperfect models. The second one is more realistic and consists of a global atmosphere model as ground truth and imperfect models that have perturbed parameters and reduced spatial resolution. In both problems we find that supermodel optimization with respect to short-term prediction error can lead to a long-term climatological behavior that is worse than that of the imperfect models. However we also show that attractor learning can remedy this problem, leading to supermodels with long-term behavior superior to the imperfect models.

Supermodels consisting of an ensemble of interacting models could be an alternative to the common approach where simulation results of separate models are collected and merged afterwards. The interactions in the supermodel need to be learned from data. The supermodel approach has been demonstrated successfully in simulation experiments with an assumed ground truth and a set of good, but imperfect models from which a supermodel is compiled with improved capability to reproduce the ground truth. In these experiments, the assumed ground truth and the imperfect models all belong to the same model class. In real applications, like climate, ecological or economical models, reality lies outside the imperfect model class. In this paper, we consider situations where the truth is much more complex than the imperfect models. We show that this can have implications for the learning strategy. Successful learning strategies are developed that should be considered, if the supermodel approach is applied to the predict the behavior or real complex systems.

I. INTRODUCTION

Synchronization is the phenomenon that coupled, oscillating systems fall into the same rhythm. Examples include clocks, singing crickets, firing neurons and applauding audiences[1]. Similar phenomena occur in multi-agent systems where synchronization mechanisms can be used to describe consensus and cooperation[2,3]. Synchronization mechanisms, by connecting models to observations, have been proposed for data assimilation[4,5]

Recently, synchronization mechanisms have been proposed as modeling improvement tools in the context of climate modeling[6,7]. The assumption is that there is a

collection of good but imperfect models. The proposal is to dynamically connect these models and so to construct one "supermodel". With sufficient connection strength, the models in the supermodel will synchronize[8,9]. The idea is that when connection strengths are learned from data, the synchronized models form an "intelligent" consensus and thereby improve their long-term climate predictions. This approach could provide an alternative to the current practice in which different climate models are run independently and their individual outcomes are combined in some form of an a posteriori ensemble average[10]. The supermodel approach is generally formulated for complex dynamical systems and is in principle widely applicable to any domain, like for example ecology, economy and biology. A promising example of supermodeling is the simulation of cancer growth where it has been demonstrated that supermodel is able to simulate qualitatively different scenarios of cancer growth which were not observed when the individual models were run separately[15,16].

To date the supermodel approach has mostly been developed in the context of weather and climate prediction. Van den Berge *et al.*[11] has performed experiments with low dimensional systems such as the Lorenz 63, the Rössler, and the Lorenz 84 oscillators[12–14]. Imperfect models were created from the assumed ground truth by parameter perturbations. The supermodels were obtained by connecting the variables dynamically, in which the connections were optimized by minimization of a short-term prediction error. As a sort of proxy for climatology, the attractors of the different models and model combinations are assessed. It was found that the optimized supermodels have attractors that are very similar to the attractors of the assumed ground truths, even when the attractors of the different imperfect models were very different.

In the original proposal, as well as in the study by Van den Berge *et al.* and other simulation studies, the supermodel connection parameters are optimized based on a short-term prediction error, resulting in a synchronised system with good long-term attractor properties[11,17]. The question is, if one can expect that this is always the case. In a more general machine learning context of chaotic attractors, Bakker *et al.*[18,19] warned that models that are optimized for short-term prediction need not automatically have good long-term (attractor) behavior as a by-product and that in general, some kind of attractor learning is needed. Similar conclusions have arisen in the context of improved representations of unresolved scales in atmosphere models by means of empirical terms in the model equations learned from data[20]. In a recent demonstration of the supermodel concept for climate modeling with real data of the tropical Pacific, a form of attractor learning has indeed been applied. The supermodel parameters were determined on the basis of the minimization of the root-mean-square difference between simulated and observed monthly mean sea surface temperature climatology statistics[21].

In this paper we want to further investigate this question. One of the model assumptions in many of the simulation studies is what we call the perfect model class scenario, in which the imperfect models and the assumed ground truth all fall in the same model class. Imperfections are due to parameter perturbations. In reality, the ground truth is often much more complex than the imperfect models, and in particular the dimensionality of the ground truth is often much larger, many of the dimensions being unresolved in the imperfect models. We call this situation, in which the ground truth model class is much more complex than the model class of the imperfect models, the imperfect model class scenario.

In this paper, we study toy problems in the imperfect model class scenario, in which we model the ground truth with observable variables and hidden variables. The imperfect models only have variables that correspond with the observable variables of the ground truth. We answer the first question negatively by showing that in such a setting, supermodels that are optimized for short-term prediction of the observables indeed can show long-term attractor properties that are much less favourable than in the earlier reported perfect model class scenario experiments, and in some sense even worse than the averaged behavior of the imperfect models.

The next question is if this is due to the supermodel concept failing in the imperfect model class scenario, or can the supermodel be improved by optimizing the connection parameters by other criteria? We will investigate this further by defining some attractor error measures. A direct form of attractor learning by optimizing these measures will be evaluated. The attractor error measures, however, are computationally expensive to evaluate. To mitigate these computational costs we make use of recently developed Bayesian optimization methods that are designed for efficient, global optimization of cost functions that are expensive to evaluate[22,23].

We first investigate and demonstrate the above mentioned phenomena in a setting where the ground truth is modeled as a 3-dimensional Lorenz 63 oscillator that is driven by another, hidden Lorenz 63 system. The imperfect models are two 3-dimensional Lorenz 63 oscillators with constant external forcing. In this constructed setting the differences between short-term prediction optimized supermodel and an attractor error optimized supermodel are quite dramatic. One could question if this result can be expected in general, or if these were atypical, due to the existence of extreme phase transitions in the attractor of the Lorenz 63 oscillator. We investigate this further in a more realistic setting with a model that is often studied in the atmospheric sciences literature. Marshall and Molteni's[24] spectral three-level, quasi-geostrophic (QG3) model simulates the winter-time atmospheric flow in the Northern Hemisphere quite realistically and produces a climatology with multiple weather regimes that are also found in observations. Meteorological fields in this model are expanded into a series of spherical harmonic functions and

are triangularly truncated at a particular total wavenumber. This truncation determines the spatial resolution of the simulations. In a perfect model class scenario with ground truth at T21 resolution and T21 imperfect models, supermodels optimized on short-term prediction have very good climatological properties[17]. In this study we will employ the QG3 model in order to see whether these results carry over to the imperfect model class scenario. In our simulation studies, the ground truth is modeled at T42, whereas the imperfect models are truncated at T21.

The paper is organized as follows. In section II we review in more detail the supermodel approach based on synchronization, and in particular in the limit of large connections and perfect synchronization. We present some earlier obtained results in the perfect model class scenario. In section III we introduce the imperfect model class scenario. We define some attractor measures that we will use in later sections. We describe the need and notion of attractor learning, and describe how we apply Bayesian optimization in the later sections. In section IV we introduce the driven Lorenz 63 oscillator as a toy problem in the imperfect model class scenario. We train supermodels by minimizing the traditional short-term prediction error as well as the attractor errors, and we discuss the results. In section V we describe the quasi-geostrophic atmosphere model that is used as a second, more realistic problem in the imperfect model class scenario. Again we train supermodels by minimizing the traditional short-term prediction error as well as the attractor errors, and we discuss the results. Finally, the results and implications of this study are discussed in section VI.

II. SYNCHRONIZED SUPERMODELS

In this section we review the supermodel approach and the findings as reported in literature[6,7,11,25]. The assumption is that there is a ground truth with an observable state $\boldsymbol{x}_{\mathrm{gt}}(t)$ that is driven by a nonlinear chaotic dynamics that is not exactly known. It is further assumed that there are M good, but imperfect models of this ground truth dynamics. These models are labeled by μ. Each of them describes the dynamics of the model state vector $\boldsymbol{x}_\mu$ according to

$$\dot{x}_\mu^i = f_\mu^i(\boldsymbol{x}_\mu) \tag{1}$$

in which i labels the vector components, and dot-notation is used for time derivatives. Then, the proposal is to combine the individual models μ into one supermodel by inserting nonnegative connections between the model equations,

$$\dot{x}_\mu^i = f_\mu^i(\boldsymbol{x}_\mu) + \sum_\nu C_{\mu\nu}^i(x_\nu^i - x_\mu^i). \tag{2}$$

With sufficient connectivity, the individual models will synchronize and reach a kind of consensus among each

other[8]. The solution of the supermodel is defined to be the average of the connected imperfect models,

$$\boldsymbol{x}_{\mathrm{s}}(t;\boldsymbol{C}) \equiv \frac{1}{M}\sum_\mu \boldsymbol{x}_\mu(t;\boldsymbol{C}). \tag{3}$$

where subscript "s" labels the supermodel state vector. The connection coefficients $\boldsymbol{C} = \{C_{\mu\nu}^i\}$ are to be inferred from a training set of historical observations. This should lead to a synchronized dynamics that converge to an attractor that is similar to the attractor of the ground truth.

One of the original proposals[7] is a scheme in which, during a training phase, additional connections with strengths $K^i > 0$ are defined that connect the model variables x_μ^i with the data from the ground truth. This causes the system not only to synchronise among each other, but also with the ground truth. The synchronisation with the ground truth during training is a form of data assimilation method called nudging[4]. Furthermore, an adaptation rule for the connections is imposed in which they follow a dynamics that depends both on the synchronisation error between the supermodel and the ground truth as well as on the synchronisation error of the imperfect models amongst each other. The result is a coupled set of differential equations for both states and parameters[7]:

$$\dot{x}_\mu^i = f_\mu^i(\boldsymbol{x}_\mu) + \sum_{\nu\neq\mu} C_{\mu\nu}^i(x_\nu^i - x_\mu^i) + K^i(x_{\mathrm{gt}}^i - x_\mu^i), \tag{4}$$

$$\dot{C}_{\mu\nu}^i = a(x_\nu^i - x_\mu^i)(x_{\mathrm{gt}}^i - \frac{1}{M}\sum_\mu x_\mu^i)$$
$$-\epsilon/(C_{\mu\nu}^i - C_{\max})^2 + \epsilon/(C_{\mu\nu}^i - \delta)^2 \tag{5}$$

In these equations, the adaptation rate a is a constant. The terms with coefficient ϵ is to dynamically constrain all connections $C_{\mu\nu}^i$ to remain in the range $(\delta, C_{\max})$. A Lyapunov function argument shows that the method converges[26].

Another approach, taken by van den Berge et al.[11] is to optimize the connections directly by numerical minimization of a short-term prediction cost function,

$$E(\boldsymbol{C}) = \frac{1}{K\Delta T}\sum_{i=1}^K \int_{t_i}^{t_i+\Delta T} |\boldsymbol{x}_{\mathrm{s}}(t;\boldsymbol{C}) - \boldsymbol{x}_{\mathrm{gt}}(t)|^2 \gamma^t dt, \tag{6}$$

which is the average of K prediction errors over a fixed time interval with length Δ, starting at K initial conditions t_i. The factor γ^t discount the errors at later times. This factor was introduced to decrease the contribution of internal error growth. With this cost function minimization approach, van den Berge et al. demonstrated the supermodel concept successfully in the context of a number of low dimensional systems such as the Lorenz 63[12], Rössler[13] and Lorenz 84[14] system. The experimental setup taken by van den Berge et al. was to define a ground truth from a model class, with a given set of ground truth model parameters. The imperfect models

were taken from the same model class, but now with considerably perturbed (imperfect) model parameters. The training yielded strong connections that indeed caused a synchronised model dynamics that was very similar to the assumed ground truth system, and notably with an attractor that was very similar to the ground truth attractor. Furthermore, in the context of the Lorenz 63 system (see equation (10) below) an experiment was performed in order to investigate whether the supermodel could predict the system response to a parameter perturbation. The result was promising: when the parameter ρ was doubled in each of the imperfect Lorenz 63 models, the optimized supermodel (without additional learning) showed a response that almost exactly mimicked the response of the assumed ground truth to a similar doubling of the corresponding ρ parameter in the assumed ground truth Lorenz 63 system.

It was demonstrated[27] that in the large connection limit, the supermodel behaves like a linear combination of the imperfect models, i.e., the states of the connected system synchronize on a single state $\boldsymbol{x}_\mathrm{s}$, of which the i component follows a dynamics which is a weighted sum of the imperfect model dynamics

$$\dot{x}_\mathrm{s}^i = \sum_\mu w_\mu^i f_\mu^i(\boldsymbol{x}_\mathrm{s}) \qquad (7)$$

with non-negative and normalized weights,

$$w_\mu^i \geq 0, \qquad (8)$$
$$\sum_\mu w_\mu^i = 1. \qquad (9)$$

The weights can be identified as unique eigenvectors of Laplacian matrices L^i that are constructed from the connections $C_{\mu\nu}^i$. Since the nonlinear differential equations of the systems studied by van den Berge *et al.* are linear in the parameters, the weighted supermodel falls in the same model class, with parameters that are weighted sums of the imperfect model parameters. Using the relation between weights and connections, it was verified that the connections that were found by van den Berge *et al.* indeed correspond to models with parameters that are close to the ground truth parameters. Also the parameter doubling experiment is straightforwardly explained by the parameter linearity[27].

Now there are two classes of supermodels, connected supermodels and weighted supermodels. The weighted supermodels can be regarded as a perfectly synchronised large connection limit of the connected supermodels. On the other hand, they can be regarded as a class of supermodels based on a simple linear combination concept, which allows for very efficient learning schemes, e.g. optimization by quadratic programming[25] and cross-pollination in time[17]. Connected supermodels are potentially richer in behaviour, due finite connection size effects and related to this, the possibility of imperfect synchronisation. Since these issues are not the subject of interest in this paper, we will restrict ourselves to weighted supermodels in the remainder of the paper and conjecture that results generalize to connected supermodels as long as their dynamics is sufficiently synchronized.

As an illustration of the result, we repeat the weighted supermodel learning experiment for the Lorenz 63 oscillator, with the same experimental set up as in van den Berge *et al.*. Both the assumed ground truth and the three imperfect models obey the Lorenz 63 equations,

$$\begin{aligned} \dot{x} &= \sigma(y - x) \\ \dot{y} &= x(\rho - z) - y \qquad (10) \\ \dot{z} &= xy - \beta z \ . \end{aligned}$$

with parameters σ, ρ and β (see Table I). The supermodel is defined by non-negative weights for each of the variables, w_μ^x, w_μ^y, w_μ^z, which all normalize to one when summed over the models, $\sum_\mu w_\mu^x = 1, \sum_\mu w_\mu^y = 1, \sum_\mu w_\mu^z = 1$. Numerical simulations are performed using a fourth order Runge Kutta scheme with a time step of $\Delta t = 0.01$. The model parameters of the assumed ground truth and the imperfect models are listed in table I. The supermodel is obtained by numerical minimization of the ΔT-time-unit-ahead prediction error, with prediction horizon $\Delta T = 0.1$

$$E(\boldsymbol{w}) = \frac{1}{K} \sum_{i=1}^{K} \sum_{t=t_i}^{t_i+\Delta T} |\boldsymbol{x}(t; \boldsymbol{w}) - \boldsymbol{x}_{gt}(t)|^2 \Delta t, \qquad (11)$$

where for each short run $t_i \to t_i + \Delta$ the weighted supermodel $\boldsymbol{x}(t; \boldsymbol{w})$ is reinitialized at $\boldsymbol{x}_{gt}(t_i)$. The supermodel itself is again a Lorenz oscillator; its parameters can be obtained by taking the inner product of imperfect model parameters and supermodel weights, $\sigma_\mathrm{s} = \sum_\mu w_\mu^x \sigma_\mu$, $\rho_\mathrm{s} = \sum_\mu w_\mu^y \rho_\mu$, $\beta_\mathrm{s} = \sum_\mu w_\mu^z \beta_\mu$. The resulting supermodel parameters are listed in Table I, from which it is seen that the parameter error is drastically reduced. The same automatically holds for the vector field error $\dot{\boldsymbol{x}}_\mathrm{s} - \dot{\boldsymbol{x}}_\mathrm{gt}$, since this is linear in the parameter error. Long-term trajectories of ground truth, imperfect models and supermodel are displayed in figure 1, from which it can be observed that the supermodel solution is almost perfect. For this model setting, the supermodel approach with standard parameter optimization by short-term prediction error minimization works very good and results in a supermodel with good attractor properties.

III. IMPERFECT MODEL CLASS SCENARIO

It is clear that the atmosphere has larger complexity and finer scales of motion than its representations by models. In mathematical terms, the real atmosphere has many more degrees of freedom, or variables, and the equations governing its evolution (if they exist) also have a much more complex structure than those of the atmospheric models. Furthermore, the state of the atmosphere is never precisely known. In the earlier work described in the previous section, the imperfect models and assumed

	σ	ρ	β
Truth	10	28	8/3
Model 1	13.25	19	3.5
Model 2	7	18	3.7
Model 3	6.5	38	1.7
SUMO	9.9	29.7	3.1

TABLE I. Lorenz 63: parameters of the assumed ground truth and the three assumed imperfect models. The supermodel (SUMO) values are obtained by taking the inner product of the imperfect model parameters with the supermodel weights w_μ^i.
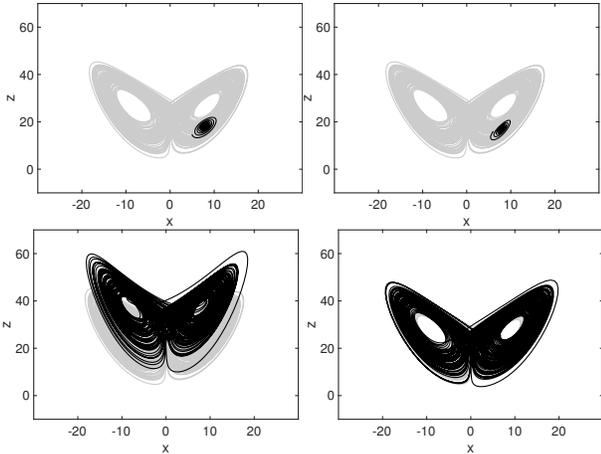

FIG. 1. Trajectories of Lorenz 63 ground truth (grey), imperfect models and supermodel (black). Upper row: Model 1, Model 2. Lower row: Model 3, Optimized weighted supermodel. Model 1 and Model 2 converge to a fixed point (in either of the wings, depending on the initial conditions)

ground truth were all from the same model class, i.e., ground truth and imperfect models have the same mathematical structure and differ only in their parameters. We call this the perfect model class scenario. In the remainder of this paper we will study the supermodel approach in the imperfect model class scenario, in which the assumed ground truth model class is more complex than the model class of the imperfect models. In particular we will assume that only a subset of variables of the ground truth is observable. The unobservable or hidden variables play the role of unresolved scales.

1. Attractor error measures

In low dimensional systems the long-term behavior of the imperfect models and supermodels can be assessed by visual inspection of the attractor. However, the visual method does not provide an objective measure and it is inadequate for systems of higher dimension. Ideally one would like to have a measure that directly measures the distance between the probability density functions describing the attractors of the model and the ground

truth.

A candidate for such a measure between probability densities is the Wasserstein metric, also known as the earth mover's distance. Intuitively, if both distributions are viewed as a unit amount of "earth", the metric is the least amount of work needed to turn one distribution into the other assuming that the cost to transporting a unit of earth from location x to y is proportional to the ground distance from x to y[28]. For systems of high dimensionality, however, this metric is impractical. Suppose a binning strategy will be taken to model the distribution of the ground truth attractor, the number of cells will be exponentially large in the number of dimensions, and an exponential large data set of observations is needed to make sure that the cells are sufficiently filled.

A more practical approach is to be less ambitious and to restrict oneself to measures based on only the state means and (co)variances of the attractors. A way to define such a measure is to approximate the distributions by multivariate Gaussians with the same means and covariances as the original distributions. The Wasserstein distance between two Gaussian distributions $\mathcal{N}_0$ and $\mathcal{N}_1$ with mean and covariances $(\boldsymbol{\mu}_0, \Sigma_0)$ and $(\boldsymbol{\mu}_1, \Sigma_1)$ respectively is given by[29,30]

$$W^2 = |\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0|^2 + \mathrm{Tr}\left(\Sigma_0 + \Sigma_1 - 2(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}\right). \tag{12}$$

which we now use as an proxy error measure between two distributions with corresponding mean and covariances. Note that this measure can be zero for non identical distributions, as long as the mean and covariance are equal. We refer to W as the Wasserstein error. To compute the Wasserstein error of a model, we collect long trajectories (not necessary of the same length) from both the ground truth and the model. From these trajectories, the mean and covariances are estimated and substituted in W.

Sometimes one is more interested in the variability of the system than in its average state[31]. In other words, it can be that one is mostly interested in the shape of the attractor and not so much in its center of gravity. In such a case, the Wasserstein error restricted to the terms containing the covariances may be more appropriate,

$$V^2 = \mathrm{Tr}\left(\Sigma_0 + \Sigma_1 - 2(\Sigma_0^{1/2}\Sigma_1\Sigma_0^{1/2})^{1/2}\right). \tag{13}$$

In high dimensional models it may be more convenient to restrict the error further to the variances $\boldsymbol{\sigma}^2$ (i.e. the diagonal terms of covariance matrix Σ), yielding the error

$$U^2 = |\boldsymbol{\sigma}_1 - \boldsymbol{\sigma}_0|^2. \tag{14}$$

We refer to the error measures W, V and U as attractor errors.

2. Attractor learning and Bayesian optimization

In the perfect model class scenario, minimizing short-term prediction error leads to supermodels that show

favourable attractor behavior, in particular improving upon the individual imperfect models. The question is if this is still the case in the imperfect model class scenario. Is it still sufficient for supermodels to be trained based on the basis of a traditional short-term prediction error and will such models always show favourable attractor behavior? We will show with counter examples that this is not the case and that the resulting supermodels can actually deteriorate the climatology compared to, e.g., the average climatology of each of the imperfect models.

The second question is if this is due to the supermodel concept failing in the imperfect model class scenario, or to the traditional learning paradigm. In a more general machine learning setting, Bakker *et al.*[18,19] has already observed that models optimized for short-term prediction do not necessarily have favourable long-term attractor behavior as by-product. When a match is desired between the observed attractor and the model attractor, some kind of "attractor learning" would be needed. Several forms of attractor learning have been proposed and investigated[18,32].

In a recent demonstration of the supermodel concept for climate modeling with real data of the tropical Pacific, a form of attractor learning has been applied. The supermodel parameters were determined on the basis of the minimization of the root-mean-square difference between simulated and observed monthly mean sea surface temperature climatology statistics[21].

In this paper, we follow a similar cost function approach, where we take the attractor errors described in the previous subsection as starting point. With the Wasserstein error W as an example, we follow the following procedure. We assume to have a sufficiently long training set of observations from the ground truth. With these we estimate once a mean and covariance of the ground truth. We further assume that we can do multiple sufficiently long runs of the supermodel and that we can estimate a mean and covariance of each of the supermodel runs. With the ground truth and supermodel statistics, a training error W_{Train} is evaluated. The training error will be a function of the supermodel parameters. In the context of this paper, these are the weights of the supermodel. Attractor learning is then the optimization of W_{Train} with respect to the supermodel weights. For each evaluation of W_{Train}, the supermodel has to be re-run.

To test the result of the supermodel training, we assume that we have an independent test set of observations from the ground truth, e.g., data from a different time window. Furthermore, we run the supermodel again for a sufficiently long time, making sure that the results are independent of the model data used to evaluate the training error, for instance by a different initialization, or by running the system for much longer time. We again gather the ground truth and supermodel statistics and evaluate W, which is now considered as test error.

Attractor learning as performed in this paper requires the evaluation of a quantity (W_{Train}) that is computed on long time supermodel simulations. Also the result is much more sensitive to phase transitions in the attractor, which may be hardly detectable in short time predictions. Global optimization methods for cost functions that are expensive to evaluate are required. Fortunately, the machine learning community has studied this problem intensively. For the optimization of computationally expensive cost functions, in particular if the parameter vector that is to be optimized is low dimensional (in our case, these will be the synchronized supermodel weights), Bayesian optimization is often proposed as the method of choice[22,23]. While probing the cost function, Bayesian optimization builds up a model of the cost function landscape, including uncertainties herein. From the predicted cost function landscape and its uncertainties, the next point to evaluate is strategically chosen. For the simulations in this paper, we used Matlab's `bayesopt` function[33], which is based on recent developments in this field[23,34,35]. We used maximally 100 cost function evaluations. We initialized the procedure applied to the supermodel learning with the M imperfect models and the supermodel with uniform weights as starting points. Furthermore we applied the `bayesopt` routine in its standard settings. Although it is recommended to follow the Bayesian optimization by a local optimizer for fine tuning, we did not pursue this further. The result that yielded the minimum of the cost function is returned and used as the supermodel solution.

The same procedure with Bayesian optimization is applied to supermodel learning based on minimization of the other attractor errors V and U, as well as on minimization of the traditional short-term prediction error E. Where applicable, the resulting supermodels are labeled as SUMO(E), SUMO(W), SUMO(V) and SUMO(U). Although more efficient or more smart procedures may exist to optimize the supermodel parameters, in particular for the traditional short-term prediction error, we applied the same procedure in all these cases since we only aim to get more insight in the two above mentioned questions, i.e., does the traditional E_{Train} minimization of the supermodel always provide good attractor properties as by-product in the imperfect model class scenario? If the answer is no, is this inherently due to the supermodel concept failing in the imperfect model class scenario or can this effect be partly remedied by other optimization procedures, in particular attractor learning? The question of computational efficiency, although partly tackled by the Bayesian optimization procedure, is deliberately not further considered in this paper.

IV. DRIVEN LORENZ63 OSCILLATOR

Our first example of the imperfect model class scenario is a low-dimensional dynamical system toy problem, the chaotically driven Lorenz 63 oscillator[36]. This example consists of a ground truth and two imperfect models.

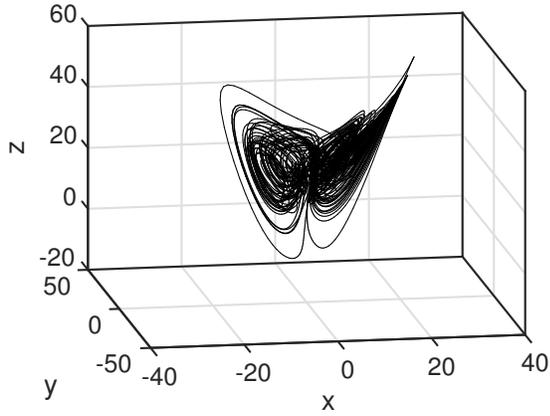The ground truth is represented by a chaotically forced

FIG. 2. Driven Lorenz 63: Trajectory of the visible variables of the assumed ground truth system.

TABLE II. Driven Lorenz 63 experiment: assumed parameter settings of the two imperfect models, Model 1 and Model 2

	σ	ρ	β	α	γ
Model 1	10	28	8/3	25	0
Model 2	6.5	38	1.6	0	10

Lorenz 63 model,

$$
\begin{aligned}
\dot{x}_v &= \sigma(y_v - x_v) + \epsilon z_h, \\
\dot{y}_v &= x_v(\rho - z_v) - y_v, \\
\dot{z}_v &= x_v y_v - \beta z_v + \delta(x_h + \eta), \\
\dot{x}_h &= \sigma(y_h - x_h), \\
\dot{y}_h &= x_h(\rho - z_h) - y_h, \\
\dot{z}_h &= x_h y_h - \beta z_h.
\end{aligned} \tag{15}
$$

in which only the v variables are visible. The hidden system h is assumed not to be directly observable. The hidden system, which itself is a chaotic system, drives the visible system v. The hidden system plays the role of unresolved scales. The parameters to model the ground truth are the following,

$$
\sigma = 10, \ \rho = 28, \ \beta = 8/3, \ \epsilon = 1, \ \delta = 5, \ \eta = 2. \tag{16}
$$

With these parameters, the shape of the trajectories of the the visible variables are somewhat similar to a perturbed Lorenz 63 "butterfly", see Fig. 2.

The two imperfect models $\mu = 1, 2$, are both represented by a Lorenz 63 system with perturbed parameters and a constant forcing,

$$
\begin{aligned}
\dot{x}_\mu &= \sigma_\mu(y_\mu - x_\mu) + \alpha_\mu \\
\dot{y}_\mu &= x_\mu(\rho_\mu - z_\mu) - y_\mu \\
\dot{z}_\mu &= x_\mu y_\mu - \beta_\mu z_\mu + \gamma_\mu
\end{aligned} \tag{17}
$$

with imperfect model parameter settings according to Table II.

Similar to the Lorenz 63 example, a (perfectly synchronized) supermodel is defined by non-negative weights for each of the variables, w_μ^x, w_μ^y, w_μ^z, which all normalize to one when summed over the two models, $\sum_\mu w_\mu^x = 1$, $\sum_\mu w_\mu^y = 1$, $\sum_\mu w_\mu^z = 1$.

Simulations are performed with a fourth order Runge Kutta scheme with step size of 0.01 time units. To generate the training set, we first run the ground truth system over a transient period of 100 time units. Then a training set of 100 time units is recorded. This period is followed by a second transient time of 100 time units and subsequently a test set is recorded of 1000 time units. For training and test set, only the visible variables (x_v, y_v, z_v) are recorded.

In the experiments with this toy problem, we consider the short-term prediction error E as defined in (11) and for the attractor errors, the Wasserstein error W, and the Wasserstein error restricted to covariances V as defined in (12) and (13). SUMO(E), SUMO(W) and SUMO(V) are optimized and results are evaluated as described in the previous section. To evaluate the attractor errors during the training phase, the supermodels are initialized each time by a state randomly drawn from a multivariate Gaussian distribution with mean and covariance estimated from the ground truth training set. Then the supermodel is first run for a transient period of 100 time units, and subsequently recorded during the next 100 time units. To obtain model data for testing, the same procedure is applied.

Trajectories of the two imperfect models and the three supermodels are displayed in Fig. 3. To generate the trajectories for these graphs, the transient period is reduced to 15 units. Data of the subsequent 100 time units is recorded and drawn in the graphs. With longer transients, the trajectories of Model 1 and SUMO(E) would collapse to a point attractor. The gray background corresponds to the first 100 time units of ground truth test data.

Test errors E, W and V of the two models and the three supermodels SUMO(E), SUMO(W) and SUMO(V) are listed in Table III. Figures in the table are based on the results of 10 repeated experiments, each with a new draw of training/test set and new optimizations. The attractor test errors of the ground truth (GT) are obtained by rerunning ground truth with different initial conditions and evaluating the attractor based on the statistics of the first and second ground truth trajectory. The initial state of the second run is randomly drawn from a 6 dimensional multivariate Gaussian, now based on the mean and covariance of all variables, including the hidden variables. Then the ground truth is run during a transient period of 100 time units, after which the visible variables are recorded during a period of another 100 time units.

To further appreciate the short-term prediction errors E of the imperfect models and supermodels, we performed an experiment in which we evaluated the training error of the ground truth with a delayed initialization of
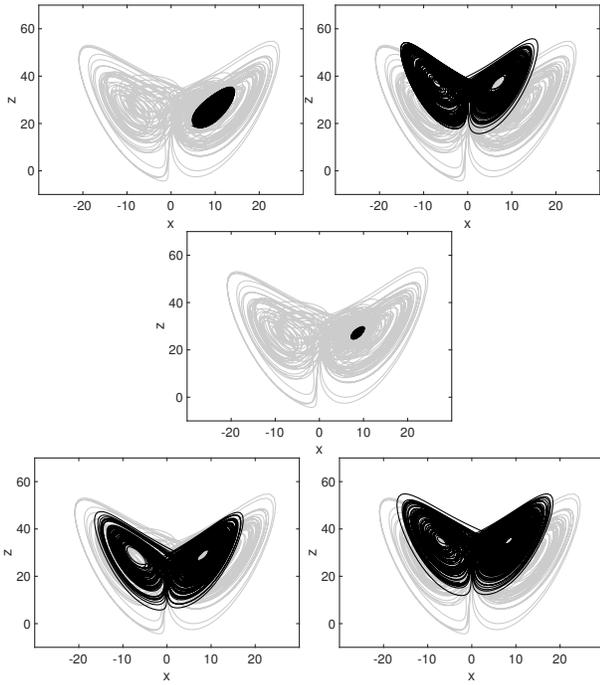
FIG. 3. Driven Lorenz 63 experiment. Top row: Model 1, Model 2. Middle row: SUMO(E). Bottom row: SUMO(W), SUMO(V). In grey: ground truth

TABLE III. Driven Lorenz 63 experiment: Errors E ($\times 10^4$), W, V for the different models and supermodels. As a reference, ground truth (GT) values are reported as well for W and V.

	Model1	Model2	SUMO(E)	SUMO(W)	SUMO(V)	GT
E	20.7(1.1)	49(4)	19.8(1.1)	21.7(1.6)	45(8)	-
W	17.9(.3)	9.9(.3)	16(3)	2.8(1.0)	8.3(.9)	1.3(.7)
V	13.9(.2)	1.4(.2)	12(3)	1.2(.5)	0.8(.4)	0.7(.4)

the hidden variables,

$$E(\tau_h) = \frac{1}{K} \sum_{i=1}^{K} \sum_{t=t_i}^{t_i+\Delta T} |\boldsymbol{x}_{\text{gt}}(t;\tau_h) - \boldsymbol{x}_{\text{gt}}(t)|^2 \Delta t, \quad (18)$$

in which $\boldsymbol{x}_{\text{gt}}(t;\tau_h)$ stands for the visible states $(x_v(t), y_v(t), z_v(t))$ of a ground truth system that is initialized for the run $t_i \rightarrow t_i + \Delta T$ in state $(x_v(t_i), y_v(t_i), z_v(t_i), x_h(t_i - \tau_h), y_h(t_i - \tau_h), z_h(t_i - \tau_h))$. As a reference, we also did the comparison with a ground truth system in which the hidden states (x_h, y_h, z_h) are initiated on their mean value estimated on the training set in which the now the hidden variables are recorded as well. Results are listed in Table IV.

Results show that SUMO(E) has the best short-term prediction performance, significantly outperforming the other models and supermodels in this measure. Model 1 is the second best model. Comparing their results in Table III with the ground truth simulation results in Table IV, one could conclude that both SUMO(E) and model 1 are reasonable short-term prediction models.

TABLE IV. Driven Lorenz 63 experiment: Errors E of ground truth with itself, when hidden state of one of the GT models is initialized with a delay τ_h, or by the average hidden state value m_h (last column)

τ_h	0	0.01	0.05	0.1	0.2	1.0	m_h
E	0(0)	1.1(.1)	5.4(.4)	10.5(.8)	18.4(1.4)	27(2)	19.8(1.4)

However, results also show that the attractor measures of these models are poor. Fig. 3 suggests that the optimized SUMO(E) converges even faster to a fixed point than Model 1, although Table III indicates that this is not always the case. The attractor performances of Model 2 are clearly better than SUMO(E). So indeed optimization of the short-term prediction error can deteriorate attractor results. Its attractor results are worse than the average imperfect model performance and actually quite close to the worst of the two imperfect model results. This can be remedied by directly optimizing the attractor errors. Fig. 3 shows that SUMO(W) has a much better attractor, in the right location and with the right shape, although still somewhat too small. SUMO(V) has an even better attractor shape (a little bit wider), although now the location is much too high. These observations can also be inferred from Table III.

The price to pay for the better attractor performance is clearly that short-term prediction errors are worse. However it is promising that for SUMO(W) these values are much better than the average prediction error of the two imperfect models, and relatively close to the best of the two imperfect models. This is, however, not the case for SUMO(V). In general one cannot expect such by-products, unless by making use of clever insight in the model structure and/or the optimization criterion, or by sheer luck.

V. QUASI-GEOSTROPHIC ATMOSPHERE MODEL

A quite realistic simulation of winter-time atmospheric flow is obtained with the three-level, quasi-geostrophic spectral model on the sphere originally constructed by Marshall and Molteni[24]. The model shows a climatology with multiple weather regimes that are also found in observations. Meteorological fields in this model are expanded into a series of spherical harmonic functions and are triangularly truncated at a particular total wavenumber. This truncation determines the spatial resolution of the simulations. Details about the partial differential equation (PDE) and how it is solved approximately in a finite state space can be found in the appendix.

In a perfect model class scenario with T21 truncated ground truth and imperfect models of the same spatial resolution but different in parameter setting, Scheven-hoven et al.[17] showed that supermodel learning based on short-term prediction (via a newly developed efficient algorithm based on cross-pollination in time) leads to a supermodel with very good climatological properties as

by-product.

Here we investigate whether this result carries over to the imperfect model class scenario. The ground truth is modeled at T42 truncation, whereas the imperfect models have a T21 truncation and different parameter setting.

In the following subsection we provide some model background information needed to understand some of the results later on.

A. Model background

The model solves the quasi-qeostrophic potential vorticity equation on the sphere at three discrete pressure levels (QG3) using a spectral method with spherical harmonics as basis functions at each pressure level. See appendix for details. The dynamical variable in QG3 is the potential vorticity (PV) in spectral coordinates at three levels, $z = 1$ (200hPa), $z = 2$ (500 hPa) and level $z = 3$ (800 hPa). Numerical solutions for this paper are obtained by applying a fourth-order Runge Kutta scheme with time steps of 1/36 day. The models in this paper are evaluated with respect to the PV in spatial coordinates $q(x, y, z, t)$ on a Gaussian grid on the Northern hemisphere with longitudes x and lattitudes y recorded at times t with intervals of 1 day.

The assumed ground truth is truncated at T42, leading to 5544 degrees of freedom. The three imperfect models are truncated at T21, with 1449 degrees of freedom. In addition to the reduction in complexity, the imperfect models differ from the ground truth in the values of a number of model parameters. These parameters, their ground truth value and perturbed values are listed in Tables V and VI. A detailed description of how these parameters enter the equations is given in the appendix. As in Schevenhoven et $al.$[17], these perturbations were created such that each of the ground truth parameters is somewhat in the middle of the perturbed imperfect model parameters. Ground truth as well as imperfect models contain PV source terms that are fitted to the mean of an observational winter climatology data set.

B. Simulations and Results

In our experiments with QG3, supermodels are defined by non-negative weights per level w_z^μ, for the levels $z = 1, 2, 3$. Again weights normalize to one when summed over the imperfect models, $\sum_\mu w_z^\mu = 1$. So with three levels and three imperfect models, we have 9 weights. Due to normalization constraints there are 6 degrees of freedom in the weights.

The supermodel approach is applied to the discretized equations that form a set of coupled ODE's, but in terms of the PDE (Eq.A.6 of the appendix) the supermodel

TABLE V. Parameters that are perturbed and their interpretation

τ_E	Time scale in days of the Ekman damping (linear damping on vorticity at lowest level)
α_1	Parameter of the land-sea mask dependent Ekman damping (more friction over land)
α_2	Parameter of the orography dependent Ekman damping (more friction over steep slopes)
τ_R	Time scale of the radiative cooling of temperature in days
τ_h	Time scale in days of the scale selective horizontal diffusion at the three levels for the smallest wavenumber
p_h	Power of the laplacian for the scale selective diffusion, the higher the more the damping is restricted to the smallest waves
h_0	Scale height of the topography in km
R_1	Rossby radius of deformation of the 200-500 hPa layer (in earth radius units)
R_2	Rossby radius of deformation of the 500-800 hPa layer (in earth radius units)

TABLE VI. Parameter settings of T42 assumed ground truth GT and three T21 imperfect models M_1, M_2, M_3

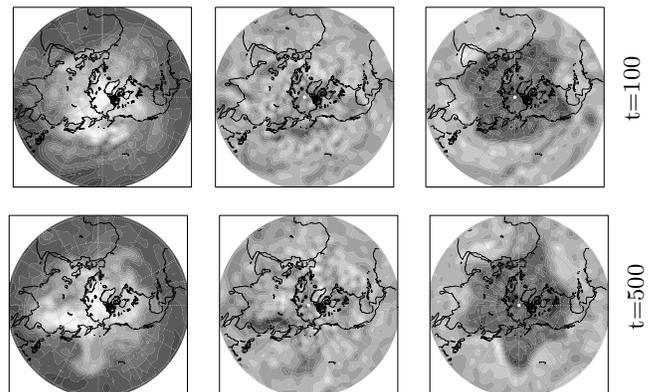|	GT	M_1	M_2	M_3
τ_E	3.0	2.0	4.0	4.0
α_1	0.5	0.2	0.8	1.0
α_2	0.5	0.2	0.3	0.1
τ_R	25	40	20	30
τ_h	3.0	5.0	4.0	2.0
p_h	4.0	4.0	2.0	3.0
h_0	3.0	9.0	5.0	2.0
R_1	0.110	0.115	0.120	0.100
R_2	0.070	0.072	0.080	0.060

FIG. 4. Potential vorticity fields at different levels and different times in assumed ground truth T42 model. From left to right: 800 hPa, 500 hPa and 200 hPa level. Top: $t = 100$. Bottom: $t = 500$. Note: different grey scales at different levels

equations are given by

$$\frac{\partial q_1^s}{\partial t} = \sum_\mu w_1^\mu \left[-\boldsymbol{v}_{\psi_1^\mu}^\mu \cdot \nabla q_1^s - D_1^\mu(\psi_1^\mu, \psi_2^\mu) + S_1^\mu \right]$$

$$\frac{\partial q_2^s}{\partial t} = \sum_\mu w_2^\mu \left[-\boldsymbol{v}_{\psi_2^\mu}^\mu \cdot \nabla q_2^s - D_2^\mu(\psi_1^\mu, \psi_2^\mu, \psi_3^\mu) + S_2^\mu \right]$$

$$\frac{\partial q_3^s}{\partial t} = \sum_\mu w_3^\mu \left[-\boldsymbol{v}_{\psi_3^\mu}^\mu \cdot \nabla q_3^s - D_3^\mu(\psi_2^\mu, \psi_3^\mu) + S_3^\mu \right] \quad (19)$$

Please consult the appendix for a detailed explanation of the symbols. Note that the velocity fields $\boldsymbol{v}_{\psi_i^\mu}^\mu$ and corresponding streamfunction fields ψ_i^μ are calculated from the supermodel PV fields q_i^s by a linear transformation that is different in each imperfect model due to perturbed parameter values.

We assume that we are mostly interested in the variability of the PV fields. Therefore we consider only optimized supermodels SUMO(E) and SUMO(U), in which E is the one-day-ahead prediction error and U the spatially averaged difference in variance in the PV fields of model and ground truth.

The training set to optimize the supermodels is obtained from the T42 ground truth model that is run for a period of 1000 days, of which the first 100 days serve as a transient, after which a training set of 900 days is recorded at intervals of 24 hours. To evaluate the results, three test sets are obtained by continuing the T42 model and recording PV fields from day 1101 to 2000, 2101 to 3000 and 3101 to 4000 respectively, again at intervals of 24 hours. In Fig. 4, the variability of the PV field $q(x, y, z, t)$ is illustrated by snapshots of the field at two different times. To generate these and other plots of spatial fields, as well as to compute the statistics of these fields that will be outlined below, the GrADS software package[37] has been used.

To evaluate the (super) model prediction error E, the potential vorticity fields (in spectral form) are projected to the wave numbers needed to obtain a T21 spectral representation. From the first day to the 899th day of the data set, these projected T42 states are used to initialize the T21 supermodel. Then the T21 model is run 899 times for 24 hours to predict the next day. The resulting predicted T21 states for the second towards the 900-th day are compared with the actual T42 states of these days, regridded to T21. The prediction error E is the root mean squared difference in T21 predicted and T42 actual state, evaluated in spatial coordinates.

To evaluate the attractor error U, the variance of PV in spatial coordinates over the period of $T = 900$ days is computed,

$$m(x, y, z) = \frac{1}{T} \sum_t q(x, y, z, t) \quad (20)$$

$$\sigma^2(x, y, z) = \frac{1}{T} \sum_t (q(x, y, z, t) - m(x, y, z))^2 \quad (21)$$

These are computed from the T42 ground truth data set and in a similar way for the models and supermodels, also
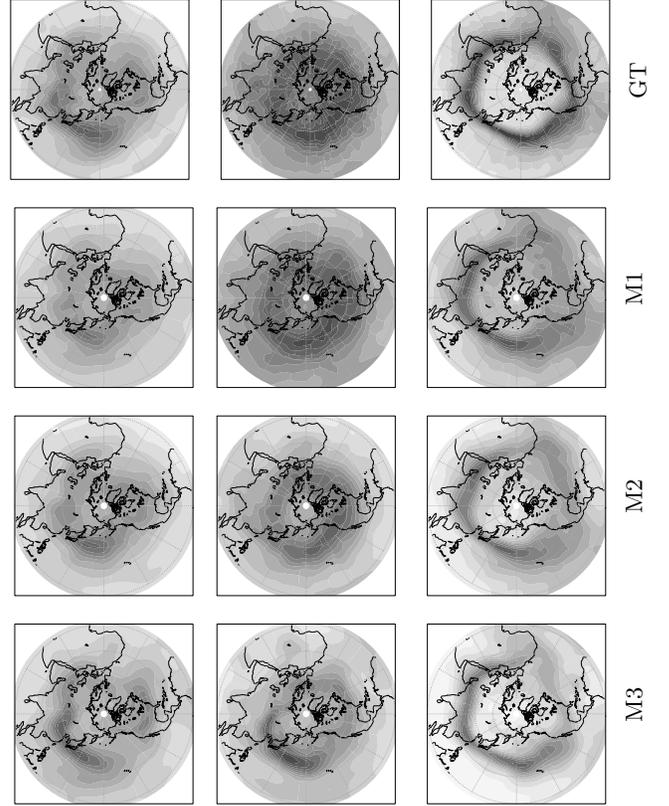

FIG. 5. Standard deviations of potential vorticity fields in ground truth T42 model and imperfect T21 models. From left to right: 800 hPa, 500 hPa and 200 hPa level. From top to bottom T42 ground truth GT, imperfect T21 models M_1, M_2, M_3. Note: different grey scales at different levels

based a simulation of sets of 900 days recorded data. To compare the statistics of T42 and T21, the T42 variances are mapped onto a the T21 grid. The attractor error U is the difference in the variances, averaged over the spatial grid and the three levels.

For illustration, level-wise standard deviation fields $\sigma^2(x, y, z)$ of the three imperfect models and ground truth are plotted in Fig. 5. In particular the sharp pentagonal region of high variability at mid-latitutes at $z = 3$ seems difficult to capture by the T21 models. Although the variability of model 1 seems to be globally the closest to the ground truth, each of the imperfect models seems to have some local features of variability in which they match the ground truth the best. The idea of the supermodel approach is to combine the models in such a way that these strengths are combined.

To train the supermodels, we used data from the T42 training set. For the training of SUMO(E), the error E of the supermodel is straightforwardly evaluated as described above. For the training of SUMO(U), at each iteration in the optimization procedure the supermodel is randomly initialized and run for a period of 1000 days of which the the last 900 days are recorded, which are

TABLE VII. Normalized errors E and U (both $\times 10^7$) for the different models and supermodels. Quantities are based on three separate (test) runs of the imperfect models M_1, M_2, M_3, and supermodels SUMO(E) and SUMO(U). In the last two columns, the values of U ($\times 10^7$) according to the non interactive posterior ensemble of the three imperfect models PE and according to the T42 ground truth GT are included for comparison. The value according to the T42 ground truth GT is the based on three independent runs of T42 compared to each other.

	M_1	M_2	M_3	SUMO(E)	SUMO(U)	PE	GT
E	159(2)	159(2)	162(2)	151(2)	154(2)	-	-
U	63(2)	88(2)	122(2)	105(5)	53(1)	69(1)	21(1)

used to evaluate the error U.

To evaluate the test error E, we used the three test sets from the ground truth. With these sets the error E of the three models and two supermodels is evaluated three times as described above. To evaluate the test error U, the three models and two supermodels are run up to day 4000, recording data from day 1101 to 2000, 2101 to 3000 and 3101 to 4000 respectively and compared with the T42 data recorded during the same time intervals. In this way the error U is evaluated three times. In the remainder of the paper, test errors are normalized, i.e., we report root mean squared errors instead of root sum squared errors. Mean and standard deviation of the normalized test errors E and U are displayed in Table VII.

As a reference, we computed the error U for the ground truth based on the difference in variance of PV in the three ground truth test sets. The result, displayed in the last column in Table VII, confirms that this error is relatively small compared to the other U-errors, which is an indication that the tuning of SUMO(U) makes sense and is not overfitting to random fluctuations of the ground truth.

Finally, we computed the error U for the non interactive posterior ensemble (PE) of the three imperfect models. The variance of the PV in the non interactive ensemble e that is needed to evaluate U is obtained by computing at each level the variance over time and models,

$$m_e(x,y,z) = \frac{1}{MT} \sum_{\mu,t} q_\mu(x,y,z,t) \qquad (22)$$

$$\sigma_e^2(x,y,z) = \frac{1}{MT} \sum_{\mu,t} (q_\mu(x,y,z,t) - m_e(x,y,z))^2 \quad (23)$$

For illustration, level-wise spatial distributions of the local attractor error u of (super) model m compared to ground truth gt

$$u(x,y,z) = |\sigma_m(x,y,z) - \sigma_{gt}(x,y,z)| \qquad (24)$$

are plotted in Fig. 6. Note that the normalized error U is the root mean squared value of u. Plotted results are based on the first test set from day 1101 to 2000 of both model and ground truth.

As a reference, we also plotted u of ground truth against ground truth (Fig. 6, last row labeled by GT). For these plots, the time periods 1101 to 2000 and 2101 to 3000 were used. The relatively small error field in the GT plots indicate that the time periods were sufficiently long for a consistent estimate of the level-wise spatial distribution of standard deviation in the potential vorticity.

Results in table Table VII show that model M_1 has indeed the smallest global attractor error U among the imperfect models. Short-term prediction errors of the three imperfect models are about the same.

With regard to question 1, we see that SUMO(E) indeed improves the short-term prediction error, as expected, but deteriorates the attractor error U. The attractor error of SUMO(E) is worse than the average attractor error of the imperfect models. With regard to question 2, we see that this can be remedied by direct optimization of the attractor error. SUMO(U) has a significantly lower attractor error U than the imperfect models, and as a (lucky) by-product, also a smaller short-term prediction error than each of the imperfect models.

If we look in more detail to the level wise spatial distributed error u, we see that SUMO(U) is not at every location the best model. For example at the 200 hPa level, the models M_2 and M_3 seem to perform a bit better in the polar area. If one would be interested in improvements in particular local areas, a straightforward method would be to put larger weights on the error in these areas in the training procedure. This is not further investigated in this paper.

VI.   DISCUSSION

Supermodeling is a modeling approach for complex dynamical systems. Its main idea is that it uses existing good, but imperfect models and combines them into one supermodel. The advantage of this approach is that it starts from existing models that were developed by domain experts, while conventional machine learning using e.g. general basis functions, starts from scratch. Although the supermodel concept is originally formulated in the context of climate science, the concept is in principle applicable to any domain involving modeling of complex dynamical systems.

The supermodel concept is originally formulated as a model consensus state by synchronization in an interactive ensemble of connected imperfect models. To get meaningful predictions, the connections in the interactive ensemble need to be optimized based on data. For the models to synchronize, large connections are needed. In the limit of large connections, perfect synchronization is obtained. The dynamics of the synchronized supermodel can be described in terms of a weighted sum of imperfect model dynamics. If we stay in this limit, the synchronized connected supermodel is described by the weighted supermodel, in which the weights are the parameters to be optimized. Since we were in primary interested in
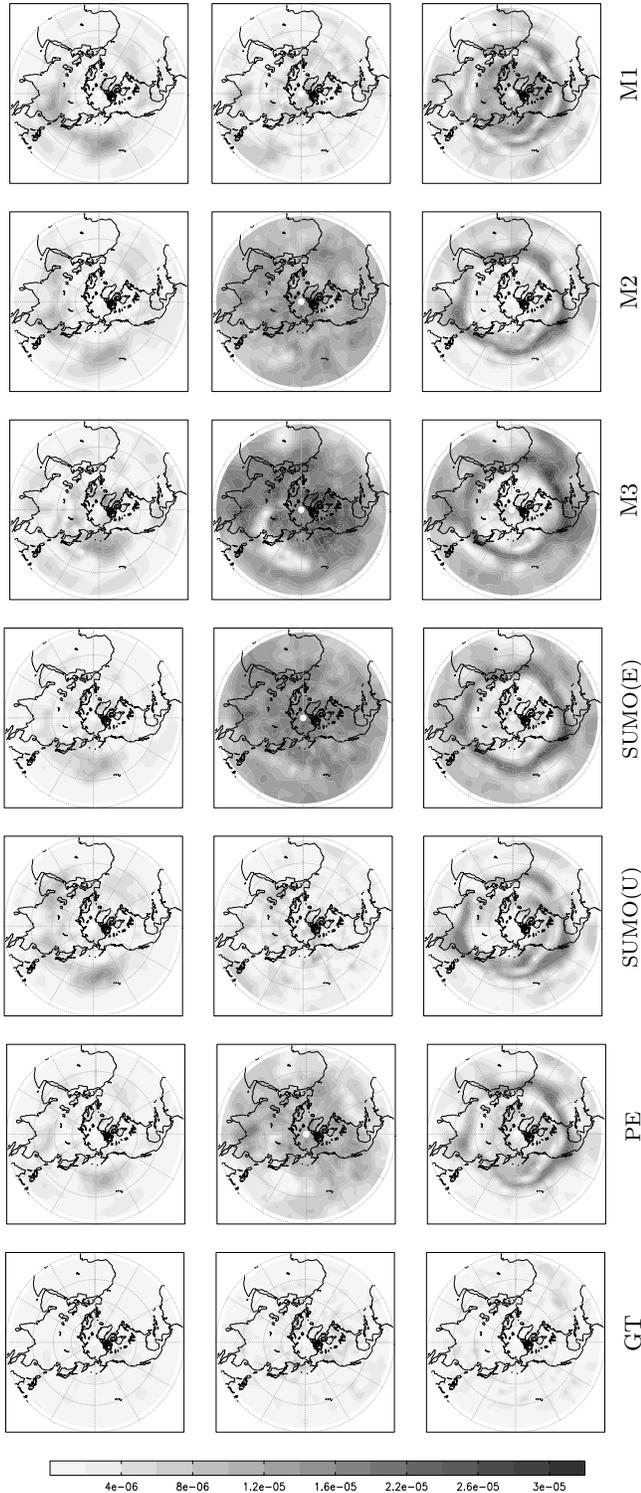
FIG. 6. Level-wise spatially distributed errors u, i.e., distributions of error in standard deviations in potential vorticity fields between various (super) model instances and assumed ground truth. From left to right: 800 hPa, 500 hPa and 200 hPa level. From top to bottom: imperfect models M_1, M_2, M_3, supermodels SUMO(E), SUMO(U), non interactive posterior ensemble of the three imperfect models PE, independent T42 ground truth run GT. Same grayscale for all levels and models

the completely synchronized regime, we have restricted ourselves in this paper to weighted supermodels and conjecture that results generalize to connected supermodels as long as their dynamics is sufficiently synchronized.

In simulations, supermodels have been studied in the perfect model class scenario, where ground truth and model differ only in their parameters, but not in model class. If in this scenario, the supermodel can exacly match the ground truth, the optimization of short-term prediction error is sufficient to tune the supermodel sufficiently well to get long-term behavior matched as a byproduct. An example of such a case is the Lorenz 63 problem described in section III.

The imperfect model class scenario where ground truth and model differ not only in their parameters, but also in model class is more realistic. In reality, it is reasonable to assume that the ground truth is more complex than the imperfect model class, in particular due to unresolved scales in the imperfect models.

Our first question was if we still can trust that in the imperfect model class scenario supermodels optimized for short-term prediction have favourable long-term attractor behavior as a by-product. If not, then the second question is if this is due to the supermodel concept that fails in the imperfect model scenario, or if this can remedied by some other kind of optimization.

We introduced attractor errors to measure the quality of long-term behavior. Our approach was to use these error measures for optimization as well, leading to a kind of attractor learning. Attractor learning is used as a somewhat brute force method to train supermodels to reproduce desired long-term behavior.

An issue with our attractor learning approach is the computational cost of training the supermodels. This is partially resolved by using Bayesian optimization, which is an efficient minimization procedure for cost functions that are expensive to evaluate.

We have first investigated our questions in a driven Lorenz 63 constructed toy problem, where the ground truth has six degrees of freedom, of which there are three observable. The imperfect models have only three degrees of freedom, corresponding with the observables. The hidden variables, which drive the visible variables in the ground truth system, serve as unresolved scales.

In the driven Lorenz 63 toy problem, we find that the first question is answered negatively. No, optimization of supermodels for short-term prediction does not guarantee favourable long-term attractor behavior as a by-product. The second question is answered positively. Yes, attractor learning can be applied, which sometimes can lead to reasonable short-term prediction performance as byproduct. However the by-product for free is not guaranteed.

The differences in the long-term behavior of the supermodels in the driven Lorenz 63 toy problem were actually quite extreme. The question is whether this observed behavior is atypical, due to e.g. the extreme sensitivity of the attractor to parameter perturbations in the Lorenz 63

system (small parameter perturbations can lead to very different attractors), or that the sub-optimal long-term behavior of supermodels trained on short-term predictions is to be expected in more realistic settings?

We therefore investigated these questions also in a more realistic setting with a model that is often studied in the atmospheric sciences literature. This model solves the quasi-geostrophic atmospheric flow equations on a sphere at three vertical levels and simulates quite realistically the winter-time atmospheric flow in the Northern Hemisphere with multiple weather regimes that are also found in observations. In a perfect model class scenario with this QG3 model run at a spatial resolution corresponding to a spectral truncation of the meteorological fields at total wavenumber 21 as ground truth and imperfect models with different parameter setting at the same T21 resolution, supermodels optimized for short-term prediction showed some very good climatological properties[17].

In our imperfect model class scenario, with a T42 truncated ground truth and T21 truncated imperfect models with different parameter setting, supermodel learning optimized for short-term prediction showed rather poor performance in the climatological behavior in terms of our attractor error. Thus the first question is again answered negatively in this more realistic setting. Also the second question is again answered positively: Yes, attractor learning can be applied and reduces the attractor error. Moreover, the supermodel optimized by attractor learning showed a reasonable prediction performance as a lucky by-product.

The first conclusion is that when the supermodel approach is applied to real data, one cannot expect that a supermodel optimized for short-term prediction automatically shows favourable long-term behavior as by-product. On the other hand, our results also show that other optimization criteria can produce supermodels which do not only perform favourably in these optimized criteria, but also may show favourable performance with respect to other criteria as a by-product.

There are still many open issues. Supermodels that perform well in other aspects than optimized for may be more trustworthy than models that only perform well in the criterium for which they are optimized. It cannot be excluded that with some clever modifications, a short-term prediction optimization strategy might mitigate the attractor problems. For example, the synchronization-based learning and the cross-pollination in time based learning might be more successful in the imperfect model scenario of QG3 of this study to produce supermodels with small attractor errors. How to find criteria and/or algorithms that lead to the best by-products, both in quantity and relevance, remains an open question. Another question is what can be learned from the performance of a set of supermodels, each of which is trained according to a different error measure U_i. Does the matrix of errors $\mathrm{SUMO}_i(U_j)$ reveal useful information about the system, the model class of imperfect models or pro-

vide guidance for the construction of more optimal supermodels?

Finally, as a by-product of this research, we were actually quite happy with the performance of the Bayesian optimization procedure. We expect that developments in efficient optimization methods like these can be very useful if supermodels are to be applied to real complex modeling.

Appendix: Derivation of the quasi-geostrophic model

Conservation of momentum on a rotating sphere and the first law of thermodynamics result in the standard filtered partial differential equations describing the temporal evolution of vorticity and temperature suitable to study the dynamics of atmospheric flow at mid-latitudes

$$\frac{\partial}{\partial t}\zeta = -\boldsymbol{v}_\psi \cdot \nabla(\zeta + f) + f_0 \frac{\partial \omega}{\partial p} \qquad (A.1)$$

$$\frac{\partial}{\partial t}\frac{\partial \Phi}{\partial p} = -\boldsymbol{v}_\psi \cdot \nabla \frac{\partial \Phi}{\partial p} - \sigma\omega \qquad (A.2)$$

Relative vorticity ζ is defined as the rotation of the horizontal wind $\nabla \times \boldsymbol{v}$ with $\boldsymbol{v} = (u,v)$, u the east-west and v the north-south component of the wind, $\nabla = \frac{1}{r}(\frac{1}{cos\phi}\frac{\partial}{\partial\lambda}, \cos\phi\frac{\partial}{\partial\mu})$, λ the geographic longitude and μ the sine of the geographic latitude ϕ, r the average radius of the Earth, $\boldsymbol{v}_\psi$ the rotational part of the wind field that can be written in terms of the streamfunction ψ as $\boldsymbol{v}_\psi = \boldsymbol{k} \times \nabla\psi$ with $\boldsymbol{k}$ the vertical unit vector. The coriolis parameter $f = 2\Omega\sin\phi$, with Ω the angular velocity of the earth, describes the contribution of the Earth's rotation to the vorticity of an air parcel at latitude ϕ and f_0 is its value at a particular reference latitude. It is also referred to as planetary vorticity. Pressure p is used as a vertical coordinate, ω is the pressure velocity which is defined as the Lagrangian rate of change of pressure with time. The vorticity equation states that the local rate of change of vorticity is due the horizontal advection of relative and planetary vorticity plus the generation of vorticity due to vertical stretching. Forcing and dissipation terms have been omitted for simplicity. Temperature is written as the pressure derivative of the geopotential Φ under the assumption of hydrostatic balance and application of the ideal gas law. Hydrostatic balance states that the pressure is equal to the weight of the atmospheric column above

$$\mathrm{d}p = -\rho g \mathrm{d}z \equiv -\rho \mathrm{d}\Phi \qquad (A.3)$$

where g denotes the gravitational acceleration, ρ the density of air and z the height. Application of the ideal gas

law $p = \rho RT$ gives a relation between temperature and geopotential

$$\frac{\partial \Phi}{\partial p} = -\frac{RT}{p} \quad (A.4)$$

with R the gas constant. Finally σ in the temperature equation denotes the vertical stability. The temperature equation states that the local rate of change of temperature is due to the horizontal advection of temperature and adiabatic heating due to vertical displacements. Combination of the vorticity and temperature equation and using the approximate linear balance equation $\nabla \Phi = f_0 \nabla \psi$ leads to a single equation for a quantity called potential vorticity (PV) that is conserved following the motion in the absence of forcing and dissipation

$$\left(\frac{\partial}{\partial t} + \boldsymbol{v}_\psi \cdot \nabla \right) \left(\zeta + f + f_0^2 \frac{\partial}{\partial p} \sigma^{-1} \frac{\partial \Phi}{\partial p} \right) = 0 \quad (A.5)$$

The quasi-geostropic model solves this partial differential equation in a finite state space with vorticity defined at discrete pressure levels 200 (level 1), 500 (level 2) and the 800 hPa level (level 3) and temperature at 650 and 350 hPa

$$\frac{\partial q_1}{\partial t} = -\boldsymbol{v}_\psi \cdot \nabla q_1 - D_1(\psi_1, \psi_2) + S_1$$
$$\frac{\partial q_2}{\partial t} = -\boldsymbol{v}_\psi \cdot \nabla q_2 - D_2(\psi_1, \psi_2, \psi_3) + S_2$$
$$\frac{\partial q_3}{\partial t} = -\boldsymbol{v}_\psi \cdot \nabla q_3 - D_3(\psi_2, \psi_3) + S_3, \quad (A.6)$$

where the index $i = 1, 2, 3$ refers to the pressure level. Here PV is defined as

$$q_1 = \nabla^2 \psi_1 - R_1^{-2}(\psi_1 - \psi_2) + f$$
$$q_2 = \nabla^2 \psi_2 + R_1^{-2}(\psi_1 - \psi_2) - R_2^{-2}(\psi_2 - \psi_3) + f$$
$$q_3 = \nabla^2 \psi_3 + R_2^{-2}(\psi_2 - \psi_3) + f(1 + \frac{h}{h_0}), \quad (A.7)$$

where R_1 (=700 km) and R_2 (=450 km) are Rossby radii of deformation appropriate to the 200-500 hPa layer and the 500-800 hPa layer, respectively and h_0 is a scale height set to 3000 m and h the height of the topography. The topography term has entered the equation through the lower boundary condition where flow over mountains lead to vertical displacement of air and the generation of vorticity through stretching. In the horizontal the equations are solved by a Galerkin projection of Eqs. (A.6) onto a basis of spherical harmonics

$$Y_{m,n}(\lambda, \mu) = P_{m,n}(\mu) e^{im\lambda} \quad (A.8)$$

where $P_{m,n}(\mu)$ denote associated Legendre polynomials of the first kind, m the zonal wavenumber and n the total wavenumber. The spherical harmonics are eigenfunctions of the Laplace operator:

$$\Delta Y_{m,n}(\lambda, \mu) = -n(n+1) Y_{m,n}(\lambda, \mu) \quad (A.9)$$

A triangular truncation of this expansion at total wavenumber 21 ($0 < n < 21, -n < m < n$) leads to a system of 1449 coupled ordinary differential equations for the 483 coefficients of the spherical harmonical functions at the three levels. For T42 the system has 5544 equations. In Eqs. (A.6), D_1, D_2, D_3 are linear operators representing the effects of Newtonian relaxation of temperature ($\mathcal{R}$), Ekman dissipation of vorticity due to linear drag on the 800 hPa wind (with drag coefficient depending on the nature of the underling surface) ($\mathcal{E}$), and horizontal diffusion of vorticity ($\mathcal{D}$)

$$-D_1 = \mathcal{R}_{12} - \mathcal{D}_1 \quad (A.10)$$
$$-D_2 = -\mathcal{R}_{12} + \mathcal{R}_{23} - \mathcal{D}_2 \quad (A.11)$$
$$-D_3 = -\mathcal{R}_{23} - \mathcal{E}_3 - \mathcal{D}_3 \quad (A.12)$$

The term

$$\mathcal{R}_{12} = \tau_R^{-1} R_1^{-2} (\psi_1 - \psi_2) \quad (A.13)$$

describes the effect of temperature relaxation between levels 1 and 2 due to radiative cooling, with a radiative time scale $\tau_R = 25$ days; the corresponding term for temperature relaxation between levels 2 and 3 is

$$\mathcal{R}_{23} = \tau_R^{-1} R_2^{-2} (\psi_2 - \psi_3) \quad (A.14)$$

The Ekman dissipation is given by

$$\mathcal{E}_3 = \boldsymbol{k} \times \nabla c_d(\lambda, \phi, h) \boldsymbol{v} \quad (A.15)$$

The drag coefficient c_d is dependent on the land-sea mask and the orographic height

$$c_d(\lambda, \phi, h) = \tau_E^{-1}[1 + \alpha_1 M(\lambda, \phi) + \alpha_2 H_d(h)] \quad (A.16)$$

with the time scale of the Ekman damping $\tau_E = 3$ days, $\alpha_1 = \alpha_2 = 0.5$; $M(\lambda, \phi)$ is the fraction of land within a grid box; and

$$H_d(h) = 1 - e^{-\frac{h}{1000}} \quad (A.17)$$

Since M and H_d vary between 0 and 1, c_d varies between $(3 \text{ days})^{-1}$ over the oceans, $(2 \text{ days})^{-1}$ over zero altitude land and about $(1.5 \text{ days})^{-1}$ over mountains higher than 2000 m. Finally, at each pressure level, the time-dependent component of PV q_i' (i.e. PV minus planetary vorticity and orographic component) is subject to a scale-selective horizontal diffusion

$$\mathcal{D}_i = c_h \nabla^{p_h} q_i' \quad (A.18)$$

where the coefficient

$$c_h = \tau_h^{-1} r^{p_h} (21 \cdot 22)^{-\frac{p_h}{2}} \quad (A.19)$$

With the power p_h set to 4 c_h is such that spherical harmonics of total wavenumber 21 are damped with time scale $\tau_h = 2$ days. The PV source terms S_i in Eqs. (A.6) are calculated from observations as the opposite of the time-mean PV tendencies obtained by inserting observed daily winter time stream function fields into Eqs. (A.6) with the PV source terms set to zero.

[1] A. Pikovsky, M. Rosenblum, and J. Kurths, *Synchronization: A universal concept in nonlinear sciences*, Cambridge Nonlinear Science Series, Vol. 12 (Cambridge University Press, 2003).

[2] R. Olfati-Saber, J. Fax, and R. Murray, "Consensus and cooperation in networked multi-agent systems," Proceedings of the IEEE **95**, 215–233 (2007).

[3] W. Yu, G. Chen, M. Cao, and J. Kurths, "Second-order consensus for multiagent systems with directed topologies and nonlinear dynamics," Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on **40**, 881–891 (2010).

[4] S. Yang, D. Baker, H. Li, K. Cordes, M. Huff, G. Nagpal, E. Okereke, J. Villafañe, E. Kalnay, and G. Duane, "Data assimilation as synchronization of truth and model: Experiments with the three-variable lorenz system," Journal of the atmospheric sciences **63**, 2340–2354 (2006).

[5] G. Duane, J. Tribbia, J. Weiss, *et al.*, "Synchronicity in predictive modelling: a new view of data assimilation," Nonlinear processes in Geophysics **13**, 601–612 (2006).

[6] G. Duane, J. Tribbia, and B. Kirtman, "Consensus on Long-Range Prediction by Adaptive Synchronization of Models," in *EGU General Assembly Conference Abstracts*, EGU General Assembly Conference Abstracts, Vol. 11, edited by D. N. Arabelos & C. C. Tscherning (2009) p. 13324.

[7] G. S. Duane, "Synchronicity from synchronized chaos," Entropy **17**, 1701–1733 (2015).

[8] L. Pecora and T. Carroll, "Synchronization in chaotic systems," Physical review letters **64**, 821–824 (1990).

[9] P. H. Hiemstra, N. Fujiwara, F. M. Selten, and J. Kurths, "Complete synchronization of chaotic atmospheric models by connecting only a subset of state space," Nonlinear Processes in Geophysics **19**, 611–621 (2012).

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

[11] L. A. van den Berge, F. M. Selten, W. Wiegerinck, and G. S. Duane, "A multi-model ensemble method that combines imperfect models through learning," Earth System Dynamics **2**, 161–177 (2011).

[12] E. Lorenz, "Deterministic nonperiodic flow," Atmos J Sci **20**, 130–141 (1963).

[13] O. Rössler, "An equation for continuous chaos," Physics Letters A **57**, 397–398 (1976).

[14] E. Lorenz, "Irregularity: a fundamental property of the atmosphere," Tellus A **36**, 98–110 (1984).

[15] W. Dzwinel, A. Kłusek, and O. Vasilyev, "Supermodeling in simulation of melanoma progression," Procedia Computer Science **80**, 999 – 1010 (2016), international Conference on Computational Science 2016, ICCS 2016, 6-8 June 2016, San Diego, California, USA.

[16] A. Kłusek, W. Dzwinel, and A. Z. Dudek, "Simulation of tumor necrosis in primary melanoma," in *Proceedings of the Summer Computer Simulation Conference*, SCSC '16 (Society for Computer Simulation International, San Diego, CA, USA, 2016) pp. 8:1–8:8.

[17] F. Schevenhoven and F. Selten, "An efficient training scheme that improves the forecast skill of a supermodel," Earth System Dynamics Discussions **2017**, 1–14 (2017).

[18] R. Bakker, J. Schouten, C. Giles, F. Takens, and C. Bleek, "Learning chaotic attractors by neural networks," Neural Computation **12**, 2355–2383 (2000).

[19] R. Bakker, *Learning to simulate and predict chaotic dynamical systems*, Ph.D. thesis, TU Delft, Delft University of Technology (2007).

[20] F. M. Selten, "A statistical closure of a low-order barotropic model," Journal of the Atmospheric Sciences **54**, 1085–1093 (1997).

[21] M.-L. Shen, N. Keenlyside, F. Selten, W. Wiegerinck, and G. S. Duane, "Dynamically combining climate models to ?supermodel? the tropical pacific," Geophysical Research Letters **43**, 359–366 (2016).

[22] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas, "Taking the human out of the loop: A review of bayesian optimization," Proceedings of the IEEE **104**, 148–175 (2016).

[23] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems* (2012) pp. 2951–2959.

[24] J. Marshall and F. Molteni, "Toward a dynamical understanding of planetary-scale flow regimes," Journal of the Atmospheric Sciences **50**, 1792–1818 (1993).

[25] W. Wiegerinck, M. Mirchev, W. Burgers, and F. Selten, "Supermodeling dynamics and learning mechanisms," in *Consensus and Synchronization in Complex Networks* (Springer, 2013) pp. 227–255.

[26] G. Duane, D. Yu, and L. Kocarev, "Identical synchronization, with translation invariance, implies parameter estimation," Physics Letters A **371**, 416 – 420 (2007).

[27] W. Wiegerinck, W. Burgers, and F. Selten, "On the limit of large couplings and weighted averaged dynamics," in *Consensus and Synchronization in Complex Networks* (Springer, 2013) pp. 257–275.

[28] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," International Journal of Computer Vision **40**, 99–121 (2000).

[29] D. Dowson and B. Landau, "The Fréchet distance between multivariate normal distributions," Journal of Multivariate Analysis **12**, 450 – 455 (1982).

[30] C. R. Givens, R. M. Shortt, *et al.*, "A class of wasserstein metrics for probability distributions." The Michigan Mathematical Journal **31**, 231–240 (1984).

[31] R. W. Katz and B. G. Brown, "Extreme events in a changing climate: variability is more important than averages," Climatic change **21**, 289–302 (1992).

[32] R. Bakker, J. C. Schouten, M.-O. Coppens, F. Takens, and C. M. van den Bleek, "Chaotic attractor learning and how to deal with nonlinear singularities," in *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000* (IEEE, 2000) pp. 466–470.

[33] The MathWorks, Natick, MA, USA, "Matlab 2016b, Statistics and Machine Learning Toolbox^{TM}," (2016).

[34] A. D. Bull, "Convergence rates of efficient global optimization algorithms," Journal of Machine Learning Research **12**, 2879–2904 (2011).

[35] M. A. Gelbart, J. Snoek, and R. P. Adams, "Bayesian optimization with unknown constraints," arXiv preprint arXiv:1403.5607 (2014).

[36] W. Wiegerinck and L. Basnarkov, "Attractor learning in interactive ensembles," in *EGU General Assembly Conference Abstracts*, Vol. 15 (2013) p. 14085.

[37] George Mason University, USA, "The Grid Analysis and Display System (GrADS)," (2016).