Linear Time-Invariant Models of a Large Cumulus Ensemble

ZHIMING KUANG[a,b]

[a] *Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts*
[b] *Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts*

(Manuscript received 19 October 2023, in final form 21 December 2023, accepted 8 January 2024)

ABSTRACT: Methods in system identification are used to obtain linear time-invariant state-space models that describe how horizontal averages of temperature and humidity of a large cumulus ensemble evolve with time under small forcing. The cumulus ensemble studied here is simulated with cloud-system-resolving models in radiative–convective equilibrium. The identified models extend steady-state linear response functions used in past studies and provide accurate descriptions of the transfer function, the noise model, and the behavior of cumulus convection when coupled with two-dimensional gravity waves. A novel procedure is developed to convert the state-space models into an interpretable form, which is used to elucidate and quantify memory in cumulus convection. The linear problem studied here serves as a useful reference point for more general efforts to obtain data-driven and interpretable parameterizations of cumulus convection.

KEYWORDS: Convective adjustment; Deep convection; Convective parameterization; Machine learning

1. Introduction

How to represent the aggregate response of a cumulus ensemble to large-scale forcing is a central issue in the dynamics of moist atmospheres. In this paper, we address this issue under two restrictions. First, we consider a cumulus ensemble sufficiently large that its behavior is approximately deterministic. While cumulus ensembles in the real atmosphere or those that parameterizations in large-scale models aim to represent are not sufficiently large to behave deterministically, it is nonetheless useful to first address the deterministic component, upon which stochastic fluctuations can then be added. Second, we consider that large-scale forcing varies around a time-invariant reference by amounts sufficiently small such that the aggregate response of the cumulus ensemble to the forcing behaves approximately linearly. This linear behavior is relevant to phenomena such as the convectively coupled waves and provides a foundation for extensions to nonlinear regimes. Much of the recent work on data-driven cumulus parameterizations through machine learning has targeted the nonlinear problem (e.g., Brenowitz and Bretherton 2018; O'Gorman and Dwyer 2018; Rasp et al. 2018; Yuval et al. 2021). It is our hope that the linear problem studied here can provide a useful reference point for such efforts.

When a cumulus ensemble responds to forcing, the effect on the large-scale environment is not instantaneous. The finite response time provides memory in convection and has been invoked as a way to preferentially damp shorter-period convectively coupled phenomena (see, e.g., Emanuel et al. 1994; Kuang 2008b). For sufficiently slowly varying large-scale forcing, the statistics of the cumulus ensemble may be assumed to be in equilibrium with the large-scale environment such that they are diagnostic functions of the latter. This will be referred to as quasi equilibrium.[1] A linear approximation of such diagnostic functions was obtained from cloud-system-resolving models (CSRM) and was referred to as linear response functions (e.g., Kuang 2010, 2012, 2018). To be more specific, hereinafter, we shall refer to them as the steady-state linear response functions. Similar work was done for jet dynamics (e.g., Hassanzadeh and Kuang 2016). However, for variations in the large-scale environment with time scales comparable to or faster than the response time scales of convection, the statistics of the cumulus ensemble may no longer be in approximate equilibrium with the large-scale environment.

In this paper, we take a time series analysis perspective and use tools developed in the field of system identification (see, e.g., Ljung 1999) to identify linear time-invariant (LTI) state-space models. These models extend the steady-state linear response functions by removing the assumption of statistical equilibrium between the cumulus ensemble and its large-scale environment. We then present a novel procedure to convert the state-space models to an interpretable form to provide some insights into memory in cumulus convection.

This paper is organized as follows. Section 2 describes the CSRM used to simulate the cumulus ensemble. Section 3 describes the state-space model framework and the experimental setup that we use. Section 4 describes the identified models and their performance, and section 5 discusses memory in cumulus convection, followed by discussion and conclusions (section 6). A number of technical details, including the description of the new procedure to convert the state-space models into an interpretable form, are given in the appendixes.

[1] In the context of cumulus convection, some confusion may arise because the quasi equilibrium of the cloud work function in the well-known Arakawa and Schubert (1974) paper refers to an equilibrium in which the convective response approximately balances the large-scale forcing. See section 5 for more discussions.

Corresponding author: Zhiming Kuang, kuang@fas.harvard.edu

DOI: 10.1175/JAS-D-23-0194.1

2. Description of the CSRM

All CSRM experiments were performed with the System for Atmospheric Modeling (SAM), version 6.7.5. An earlier version of this model was described in Khairoutdinov and Randall (2003). The model solves the anelastic equations of motion. The prognostic thermodynamic variables are liquid water static energy, total nonprecipitating water, and total precipitating water. We use a bulk microphysics scheme and a simple Smagorinsky-type 1.5-order scheme to parameterize the effect of subgrid-scale turbulence. Surface latent and sensible heat fluxes are computed using bulk aerodynamic formula with a constant 10-m exchange coefficient of 1×10^{-3} and a constant surface wind speed of 5 m s^{-1}. Surface momentum fluxes are computed with the Monin–Obukhov similarity theory.

The domain is 128 km × 128 km in size in the horizontal with a 4-km resolution and doubly periodic lateral boundary conditions. There are 28 vertical layers that extend from the surface to 32 km, the top third of the domain being a wave-absorbing layer. The top two layers in this model are restored to prescribed values with a 1-h time scale and will be excluded from our forcing and analysis. The relatively coarse horizontal and vertical resolutions match those used in the Superparameterized Community Atmosphere Model (SPCAM), which has been shown to produce the convectively coupled waves, convective self-aggregation, and the Madden–Julian oscillation (e.g., Khairoutdinov et al. 2008; Arnold and Randall 2015), and were chosen to both reduce the computational cost and to make the results directly relevant to other studies using the SPCAM.

All experiments are over an ocean surface with fixed sea surface temperature of 29°C. Idealized radiation is prescribed following Pauluis and Garner (2006): a cooling rate of 1.5 K day^{-1} is used when temperature is greater than 207.5 K; otherwise, Newtonian relaxation to 200 K is used with a time scale of 5 days. The mean state is that of a radiative–convective equilibrium (RCE) and no mean wind shear; the latter condition is ensured by damping the horizontally averaged horizontal winds with a time scale of 15 min. The approach described in this paper, however, can be applied to any mean state.

3. System identification

a. Generation of input–output data

A key step in identifying LTI models is to obtain sequences of input and output data with adequate excitation across different modes and frequencies, sufficient signal-to-noise ratio (SNR), and minimal nonlinear distortion.

Define $\mathbf{f}_t$ as the input vector of size $m \times 1$ such that that $\mathbf{f}_t/\Delta$ is the anomalous temperature T and specific humidity q tendencies that we impose uniformly in time over the time interval $[t\Delta, (t + 1)\Delta]$, where Δ is the sampling interval and t is the time index. The imposed tendencies are in addition to the reference RCE forcing. They are horizontally uniform over the CSRM domain and represent tendencies due to large-scale dynamics. Define $\mathbf{y}_t$ as the $m \times 1$ output vector that describes deviations of the instantaneous CSRM horizontal mean T and q profiles at time $t\Delta$ from their respective reference profiles. The sizes of the input and output vectors are chosen to be the same in this study but need not be so in general. We set Δ to 15 min to match the global model time step often used in SPCAM. As described in section 2, the top two levels of the model are nudged to reference values with a time scale of 1 h, they therefore do not vary freely and are excluded from the inputs and outputs. In addition, only humidity and humidity forcing of the lowest 14 layers (below ~350 hPa) are included in the inputs and outputs; with the idealized radiation, humidity variations at higher levels have minimal effects on convection. With these choices, inputs and outputs are both 40×1 vectors; that is, $m = 40$. While horizontal averages of horizontal winds and their forcing can be included in the inputs and outputs as well, they are excluded in this study of a zero mean shear case.

To enhance the SNR, we ran 2048 copies of the CSRM with slightly different initial conditions in the RCE setup described in section 2 for 200 days without additional forcing (i.e., zero input) to reach a statistically steady state. These copies are thus different realizations of the cumulus ensemble in a 128 km × 128 km CSRM and will be referred to as the ensemble members. All the ensemble members were run for another 1000 days using identical time sequences of $\mathbf{f}_t$. The 1000-day sequences of $\mathbf{f}_t$ were made by first generating 200-day input sequences using randomly phased multisines and then repeating them five times. Appendix A describes the details of input sequence generation and how the forcing amplitudes were adjusted to balance SNR and nonlinear distortion.

The first of the five 200-day periods was discarded as it was affected by transients (the initial adjustments) that lasted for ~50 days (not shown). The remaining four 200-day periods, having identical inputs and unaffected by transients, were averaged. This gives the input and output sequences $\mathbf{f}_t$ and $\mathbf{y}_t$ from one experiment. Variations among the four 200-day periods provide an estimate of the stochastic noise.

We ran a total of eight experiments. Each experiment differed only in the random number sequence used to generate the random phases of the multisines (see appendix A). Six of the experiments were combined to produce the dataset upon which the system identification was performed, further reducing the stochastic noise and the stochastic nonlinear distortion. The other two experiments were used for validation. The above procedure largely follows the recommendations in Pintelon and Schoukens (2012), which contains more discussions on the input–output data generation.

b. State-space model

There is an extensive literature on the identification of LTI systems from input and output data and the readers are referred to Ljung (1999) for a systematic exposition. In the following, we briefly describe the framework, which also serves to establish the notation.

We will describe the behavior of the cumulus ensemble as a discrete LTI system in the state-space form, which has

become the dominant way of representing linear dynamical systems since the celebrated work of Kalman (1960):

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{f}_t + \boldsymbol{\sigma}_t \quad \text{and}$$
$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \boldsymbol{\mu}_t. \tag{1}$$

Here, $\mathbf{y}_t$ and $\mathbf{f}_t$ are the input and output vectors defined in section 3a, and $\mathbf{x}_t$ is the $n \times 1$, yet unknown, internal state vector that captures the state of the system at time $t\Delta$ in an n-dimensional hidden or latent space, with n being the order of the model; $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ are time-invariant matrices that capture the system dynamics; $\boldsymbol{\sigma}_t$ and $\boldsymbol{\mu}_t$ are (possibly colored) Gaussian noises from unrepresented processes during the time interval $[t\Delta, (t + 1)\Delta]$ and measurement noises at time $t\Delta$, respectively. The Gaussian assumption is justified here by the central limit theorem, as we are concerned with the average of a large ensemble of independent and statistically identical cumulus fields. All effects of $\mathbf{f}$ on $\mathbf{y}$ are assumed to go through $\mathbf{x}$. With the Gaussian assumption for the process and measurement noises, Eq. (1) can be rewritten in the innovation form [see section 4.3 of Ljung (1999) for details]:

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{f}_t + \mathbf{K}\mathbf{e}_t \quad \text{and}$$
$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{e}_t, \tag{2}$$

where $\mathbf{K}$ is the Kalman gain matrix and $\mathbf{e}$ is a Gaussian white noise term with covariance matrix $\mathbf{R}$, commonly referred to as the innovation. The matrices $\mathbf{K}$ and $\mathbf{R}$ will be directly estimated from the input–output data and are connected to the covariance and cross-covariance matrices of the process and measurement noises and the matrices $\mathbf{A}$ and $\mathbf{C}$.

c. *Parameter estimation*

With an input–output dataset and a user specified model order, n (size of the internal state vector $\mathbf{x}$), parameters $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{K}$, and $\mathbf{R}$, and the initial state $\mathbf{x}_0$ of the state-space model [Eq. (2)] can be estimated by minimization with well-established methods (Ljung 1999). One can choose to minimize the simulation error or the one-step-ahead prediction error (hereinafter referred to simply as prediction error).

The prediction error is calculated by driving the model with the inputs and past measured outputs using the Kalman gain:

$$\tilde{\mathbf{x}}_{t+1} = \mathbf{A}\tilde{\mathbf{x}}_t + \mathbf{B}\mathbf{f}_t + \mathbf{K}\tilde{\mathbf{e}}_t,$$
$$\tilde{\mathbf{y}}_t = \mathbf{C}\tilde{\mathbf{x}}_t, \quad \text{and}$$
$$\tilde{\mathbf{e}}_t = \mathbf{y}_t - \tilde{\mathbf{y}}_t, \tag{3}$$

where model results are indicated with a tilde. The simulation error is the error when Kalman gain is set to zero in Eq. (3). The weighted L_2 norm of the residual $\left\|\mathbf{w} * \tilde{\mathbf{e}}_t\right\|_2$, averaged over t, is minimized in the respective cases, where * denotes elementwise multiplication of two vectors and $\mathbf{w}$ is a weighting vector with elements w_i, defined as

$$w_i = \begin{cases} \sqrt{\Delta m_i} \text{ if } y_i \text{ is temperature} \\ \dfrac{L}{c_p}\sqrt{\Delta m_i} \text{ if } y_i \text{ is humidity} \end{cases}, \tag{4}$$

in which L is the latent heat of vaporization of water, c_p is the specific heat of air, and Δm_i is the mass of layer i. As compared with minimizing the simulation error, minimizing the prediction error is equivalent to inversely weighting the model error by noise such that noisy components and frequencies are weighted less [see chapter 12 of Ljung (1999) for more details].

We estimate the parameters with the subspace method, an efficient noniterative method implemented in the *n4sid* function of the commercially available MATLAB System Identification toolbox. For a technical description of the subspace algorithm, see Ljung (1999), section 10.6. The identified model could in principle be further improved using iterative methods. This is not done here, as the subspace solution appears to be sufficiently accurate, and the iterative methods take an impractically long time. With the subspace method, there are several additional user (or hyperparameter) choices to make. These are described in more detail in appendix B.

4. Identified state-space models

a. *Validation errors*

All of the state-space models that we have identified using procedures described in section 3 are stable; that is, the eigenvalues of their matrix $\mathbf{A}$ all have modulus less than 1. These models are then applied to the validation set in the simulation mode, the mode that we intend to use the models in. For each identified model, after averaging the weighted L_2 norm of the residual $\tilde{\mathbf{e}}_t$ and that of $\mathbf{y}_t$ over t, we take their ratio and show it as a percentage in Fig. 1, with Fig. 1a being a broad overview and Figs. 1b and 1c being close-ups on the small and large model order portions, respectively. The different hyperparameter choices represented by the different symbols in Fig. 1 are described in appendix B. Since simulation error stabilizes beyond the model order of ~150, we shall use $n = 160$ with canonical variate analysis (CVA) weighting (Larimore 1990) and prediction focus (described in appendix B) as our reference model and will focus on it in the remainder of this section. It is also worth noting that the simulation error continues to decrease significantly as n increases beyond 40. Models with n greater than 40, the size of the output vectors, include effects from past outputs. This memory effect will be discussed in section 5.

b. *Transfer function*

The transfer function of the state-space model is a frequency-dependent $m \times m$ matrix that describes how outputs depend on inputs as a function of frequency ω:

$$\mathbf{G}(\omega) = \mathbf{C}(\mathbf{I}e^{i\omega\Delta} - \mathbf{A})^{-1}\mathbf{B}, \tag{5}$$

where $\mathbf{I}$ is the identity matrix. In and only in this expression and Eq. (22) later in the paper, i denotes the imaginary unit. While ω in Eq. (5) can be complex valued, we will restrict ourselves to real frequencies. To provide a broad overview of the transfer function and its accuracy, we show in Fig. 2 amplitudes of the response and the model error as functions of frequency for one of the experiments in the validation set; results from the other experiment are very similar. Note that input forcing

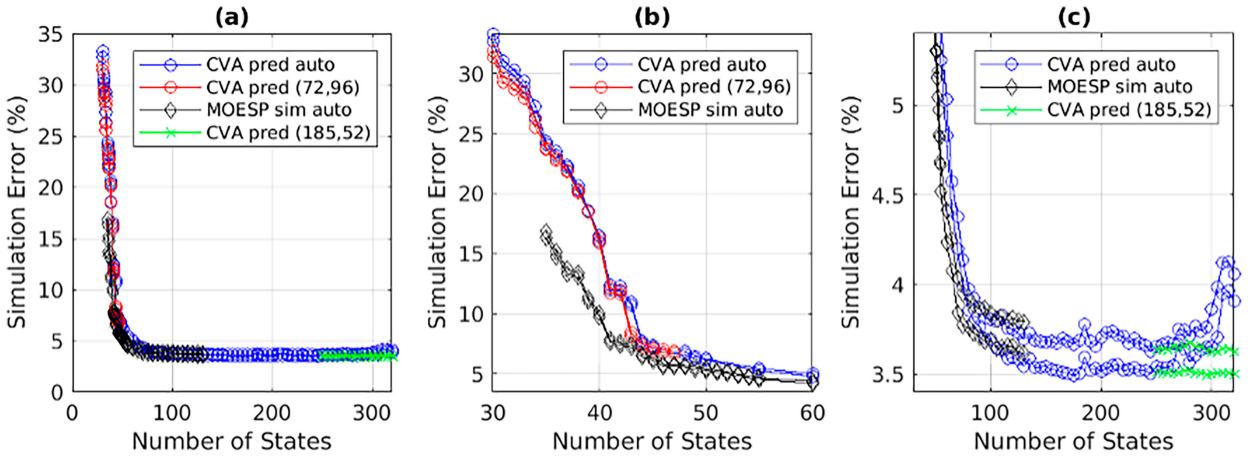

FIG. 1. Percentage errors when identified state-space models with a range of model orders are used without the Kalman gain to simulate the outputs given the inputs from the two validation experiments. The simulation errors are plotted as functions of the number of states used in the models (i.e., model order). (a) An overview, and details on the (b) low- and (c) high-model-order ends. Blue circles are for canonical variate analysis (CVA) weighting (Larimore 1990) and prediction focus with the number of past and future inputs and outputs determined by the Akaike information criterion (AIC), and they are labeled as "CVA pred auto." Red circles and green crosses are the same as blue circles, but they use 96 past and 72 future inputs and outputs [labeled as "CVA pred (72, 96)"] and 52 past and 185 future inputs and outputs [labeled as "CVA pred (185, 52)"], respectively. Black diamonds are for multivariate output-error state space (MOESP) weighting (Verhaegen 1994) and simulation focus with the number of past and future inputs and outputs determined by the AIC, and they are labeled as "MOESP sim auto." See the main text and appendix B for more details.

amplitude is a smooth and monotonic function of frequency [see Eq. (A2) in appendix A].

The blue curve in Fig. 2a shows the weighted average of the power spectra of the individual components of **y**, using the square of the weighting shown in Eq. (4). This will be referred to as the signal. The same quantity for the stochastic noise in **y** is shown in orange. As noted in section 3, the stochastic noise was estimated from variations among the four 200-day periods with identical forcing. To compute the model error (red curve in Fig. 2a), we apply the identified reference model with $n = 160$ in the simulation mode to the validation experiment, compute the power spectra of the errors in individual components of **y**, and then compute their weighted average as done for the signal.

Figure 2a shows that the simulation error is at the level of the stochastic noise for all frequencies. Figures 2b and 2c show the same comparison for the power spectra of the column-averaged moist static energy (MSE), in temperature unit, and net precipitation (precipitation minus evaporation) measured in millimeters per day. The time series of the net precipitation is computed from **y** and **f** based on conservation of water; contributions from condensed water are small and neglected. These results indicate that, while the model is identified by minimizing the weighted L_2 norm of the residue, it also performs well in other metrics. Figure 2d shows the variance of T and q as functions of pressure for additional context.

Despite being forced more strongly at higher frequencies (see appendix A), the weighted average of the power of **y** (Fig. 2a) and the column-averaged MSE (Fig. 2b) both have considerably greater power at lower frequencies, with the SNR (in terms of power) ranging from several hundreds to 1000 or more. For net precipitation, there is a broad peak in the noise spectrum at periods of several hours, while the signal spectrum is relatively flat for frequencies lower than a few hours. The stochastic noise is nonwhite in all three metrics in Figs. 2a–c, and there is a sharper peak around the period of ~6 h, indicating the presence of some internal stochastic processes with that time scale. This peak is also present in the signal in both the weighted average of the power spectra of **y** and the power spectrum of the net precipitation, indicating the presence of resonance because the prescribed forcing is smooth in frequency (see appendix A).

c. *Noise model*

Parameter matrices **A**, **C**, **K**, and **R** provide a model for the stochastic noise in the system. To evaluate this noise model, we use deviations of the 2048 individual ensemble members from the ensemble mean to provide an independent and explicit estimate of the stochastic noises in an individual ensemble member. Covariance matrices of these stochastic noises for a range of lags are output from the CSRM and compared with those computed from the noise model based on parameter matrices **A**, **C**, **K**, and **R**. The formulas for such computation are presented in appendix C. The comparisons show broad agreement and two examples, with 0 lag and 3-h lag, are given in Fig. 3.

The lagged covariance from the noise model therefore can be useful to stochastic modeling efforts such as Neelin et al. (2008) and Palmer (2019) with the caveat that noises in individual ensemble members may not be Gaussian. Further studies are needed to fully characterize the noise distributions in individual ensemble members.

d. *Steady-state linear response function matrix*

The steady-state linear response function matrix describes the convective tendencies for a T and q profile that is steady

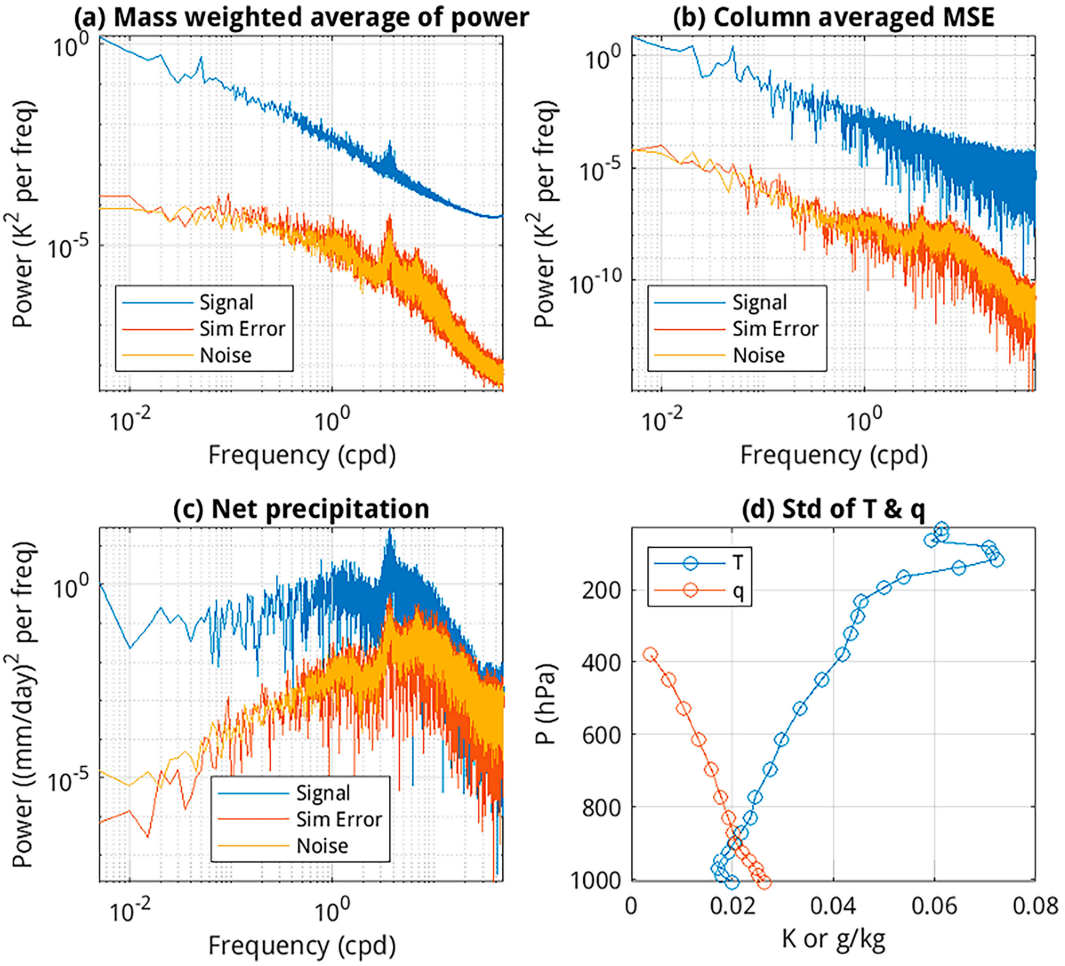

FIG. 2. (a) Mass-weighted average of the power spectra of the output signal **y** (blue) and its stochastic noise (orange) from one of the experiments in the validation set, along with the simulation error using the identified state-space model with model order $n = 160$. (b) As in (a), but for column-averaged moist static energy (MSE). (c) As in (a), but for net precipitation (precipitation minus evaporation). (d) Standard deviations of temperature (blue) and specific humidity (red) in the output signal **y** as functions of pressure.

in time and can be computed from the state-space model parameters as

$$\mathbf{M} = -[\mathbf{C}(\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}]^{-1}. \tag{6}$$

The matrix **M** is found to have a single positive eigenvalue of ~30 day^{-1}. Positive eigenvalues of the steady-state linear response function were also found in Kuang (2010, 2018) and were attributed to estimation error. Such an explanation can be ruled out in this study given its high SNR. We confirm the accuracy of the positive eigenvalue using additional CSRM experiments with time-invariant forcing that aligns with the eigenvector associated with this positive eigenvalue (top panels in Fig. 4). Further efforts, described in appendix D, were made to polish the eigenvector to enhance clarity, the results from which are shown in the bottom panels of Fig. 4.

Figure 4 shows that warming and moistening the boundary layer and cooling and moistening the free troposphere (with horizontally uniform forcing) in accordance with the eigenvector of **M** with the positive eigenvalue make the boundary layer colder and drier and the free troposphere warmer and drier, all measured in terms of horizontal averages. This seemingly counterintuitive result will be interpreted physically using the joint probability distribution function (PDF) as additional information in section 6.

e. *Coupling with two-dimensional gravity waves*

To further evaluate the identified models, we couple the state-space models and two-dimensional (2D) linear gravity waves with a single horizontal wavenumber k. As described in Kuang (2008a), for linear gravity waves of a single horizontal wavenumber, hydrostatic balance, the horizontal momentum equation, and the continuity equation can be combined to give an equation [their Eq. (7)] that relates time tendencies of the vertical velocity profile to the anomalous temperature/buoyancy profile. For brevity, we will not repeat the derivation here.

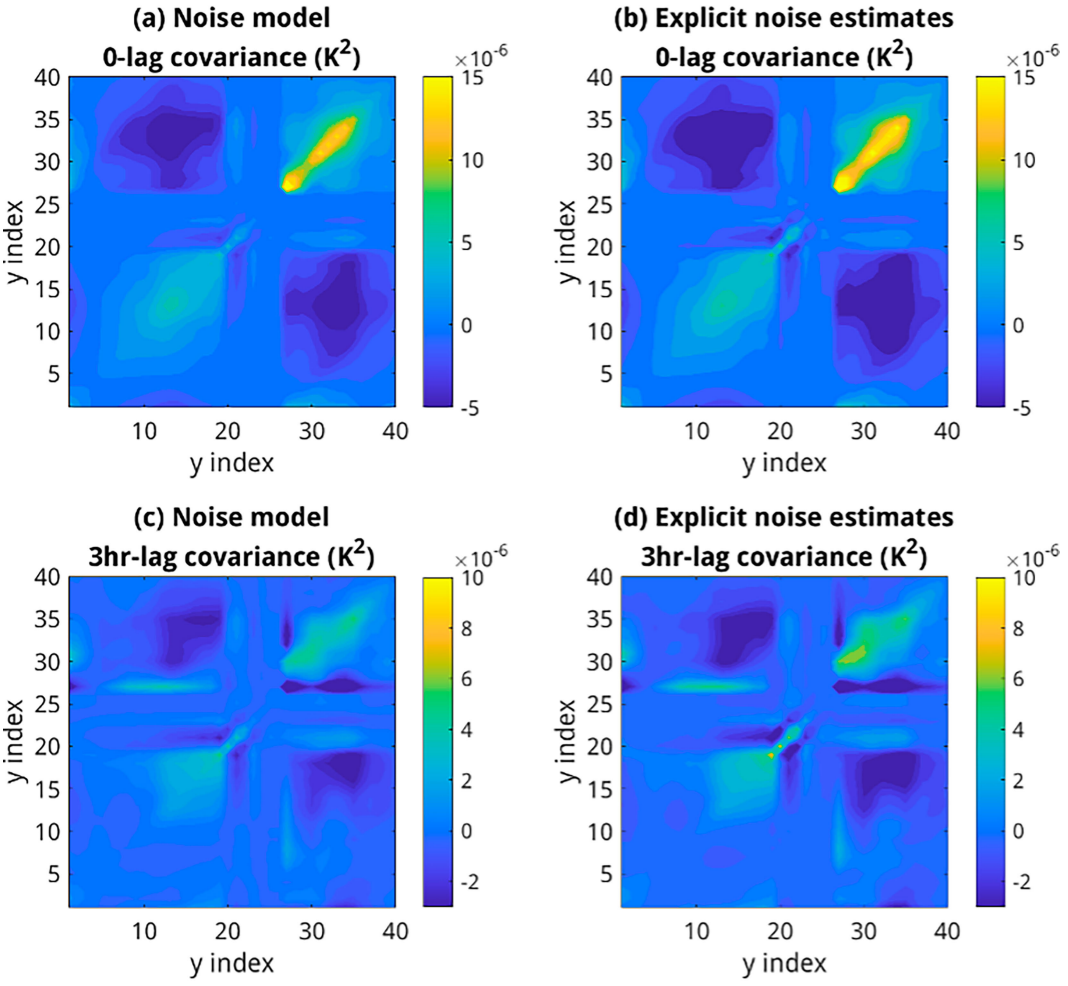

FIG. 3. Comparison of the (a),(b) 0-lag and (c),(d) 3-h-lag noise covariance matrices from the (left) identified noise model and (right) explicit estimates. For the comparison, the explicit estimates of the covariance matrices are divided by 2048, the size of the ensemble, and the noise-model results are multiplied by 4 since the noise model is estimated using the averages of four 200-day periods with identical inputs.

Combining this equation and the state-space model using temperature and moisture advection due to the anomalous vertical velocity as the inputs, we have

$$\begin{pmatrix} \mathbf{x}_{t+1} \\ \mathbf{w}_{t+1} \end{pmatrix} = \begin{pmatrix} \mathbf{A} & \mathbf{BF} \\ k^2\mathbf{EC} & \mathbf{D} \end{pmatrix} \begin{pmatrix} \mathbf{x}_t \\ \mathbf{w}_t \end{pmatrix} + \begin{pmatrix} \mathbf{AK} \\ k^2\mathbf{E} \end{pmatrix} \mathbf{e}_t, \tag{7}$$

where $\mathbf{Fw}_t$ represents the forcing due to the advection of the reference temperature and moisture profiles by vertical velocity $\mathbf{w}_t$, $k^2\mathbf{E}$ describes how $\mathbf{y}$ (anomalous temperature and moisture profiles) affects the vertical velocity profile $\mathbf{w}$ (note that $\mathbf{y}_t = \mathbf{Cx}_t + \mathbf{e}_t$), and $\mathbf{D}$ represents momentum persistence, here including only the effect of specified Rayleigh damping as in Kuang (2008a). The other variables are the same as those in Eq. (2). This system is similar to that of Kuang (2012, 2018) except written in a discrete form, using state vector $\mathbf{x}$ instead of output $\mathbf{y}$, and with the addition of stochastic noise $\mathbf{e}$. A wave-radiating upper boundary condition is imposed at ~25 hPa. With our idealized radiation, water vapor variations above 350 hPa have little effect on convection or the large-scale waves and are neglected.

Figure 5 compares the simulated rain rates when the 2048-member CSRM ensemble and the reference state-space model are coupled to 2D gravity waves with a horizontal wavelength of 2000 km. Coupling of the state-space model with 2D gravity waves was done through Eq. (7). Three realizations of Eq. (7) using different random seeds are shown to indicate the range of behaviors. With a Rayleigh damping time of 0.23 days, coupled simulations with either the CSRM or the state-space model produce growing convectively coupled waves (Fig. 5a). With a Rayleigh damping time of 0.20 day, no growing waves are seen in either case (Fig. 5b). The stability boundary in terms of the Rayleigh damping rate therefore differs by no more than 15% between the CSRM and the state-space model. Figure 6 shows a similar result for the horizontal wavelength of 5000 km.

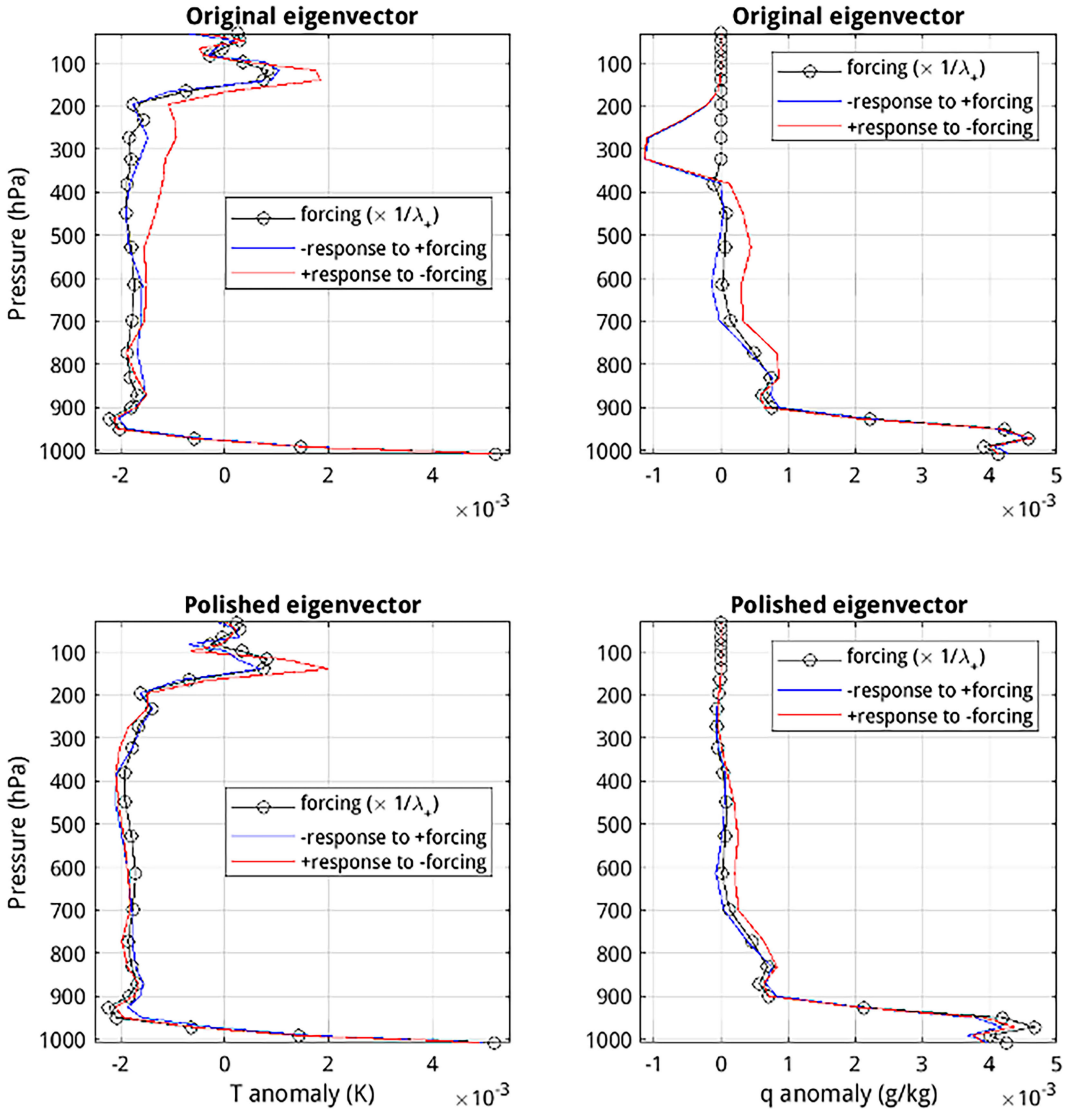

FIG. 4. (top left) Steady-state temperature and (top right) specific humidity responses of the CSRM ensemble to forcing that aligns with the eigenvector of **M** that is associated with the positive eigenvalue. The forcing was divided by the eigenvalue (~30 day^{-1}) and is shown as black circles. The signs of the responses to this forcing have been reversed (blue lines) for closer comparison. Red lines show the responses when a forcing with reversed signs is used. (bottom) As in the top panels, but with forcing that aligns with the polished eigenvector. The procedure to polish the eigenvector is described in appendix D.

Blue circles in Fig. 7a give an overview of the stability of the convectively coupled waves represented by the system of Eq. (7) without stochastic noise or Rayleigh damping, showing the growth rate as a function of the horizontal wavenumber. One notable result is that the highest growth rate occurs at a horizontal wavelength of about 1700 km. This is considerably shorter than the wavelength at which the normalized spectral power for Kelvin waves maximizes in observations and in SPCAM simulations (e.g., Wheeler and Kiladis 1999; Khairoutdinov et al. 2008). While past studies have invoked convection memory due to the finite time lag in convective response to large-scale forcing to shift the peak in growth rates to lower wavenumbers (e.g., Emanuel et al. 1994; Kuang 2008b), such effects are already accounted for in the state-space model. Therefore, other, yet to be explored, processes are needed to account for the difference.

Results in this section establish that the reference state-space model with $n = 160$ gives a sufficiently accurate representation of how horizontal averages of temperature and humidity of the CSRM ensemble evolve with time under small forcing and can be used to further our understanding of convectively coupled phenomena. We now turn our attention to models with lower orders to obtain some insights into the importance and the nature of memory in cumulus convection.

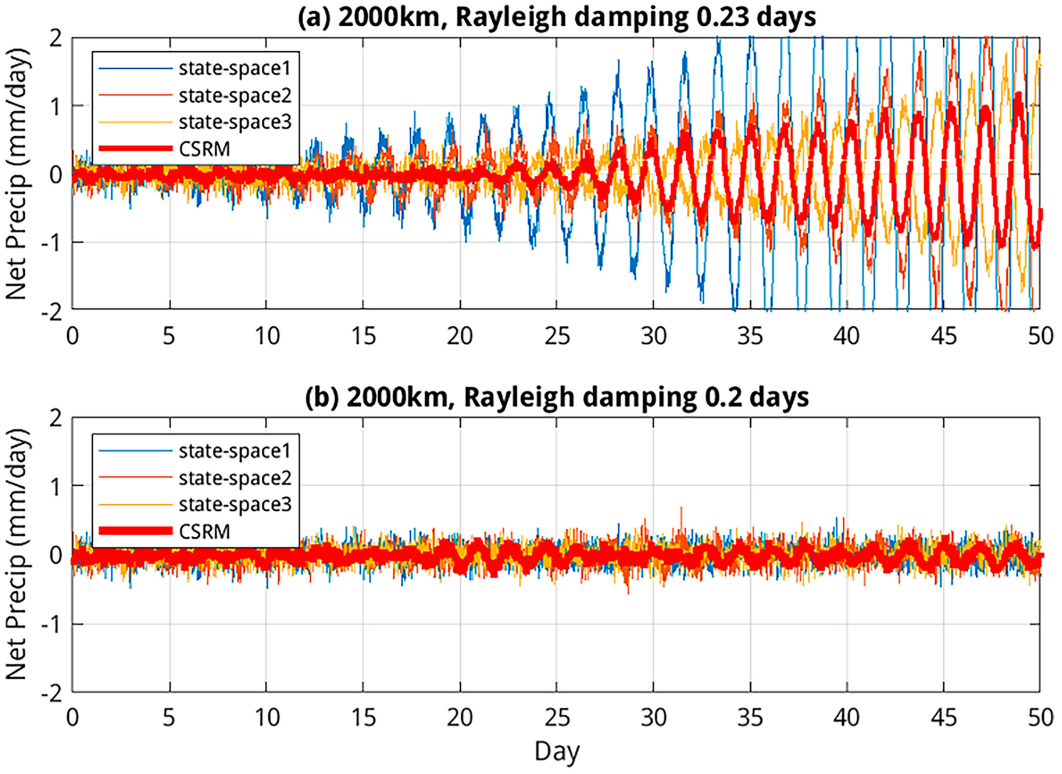

FIG. 5. Time series of net precipitation (precipitation minus evaporation) after the CSRM ensemble (thick line) and the reference state-space model (thin lines) are coupled to 2D gravity waves of 2000-km horizontal wavelength and Rayleigh damping of (a) 0.23 and (b) 0.2 days. The three thin lines are for three different random noise sequences used in the state-space model.

5. Conversion of state-space models for interpretations in terms of convection memory

A state-space model identified in a multivariate case is generally considered a "black box." While canonical forms can be constructed in the univariate case to assist with interpretation, canonical forms for multivariate cases (and the corresponding input–output models) are nonunique (Luenberger 1967). We have designed a procedure to facilitate the interpretation of the identified state-space models in terms of convection memory. Before we start, however, it is important to note that the equations solved by the CSRM itself are autonomous. Memory only arises because we have reduced the prognostic variables of the CSRM to horizontally averaged profiles of T and q—our $\mathbf{y}$; if this reduction removes information needed to determine future values of $\mathbf{y}$, past values of $\mathbf{y}$ are needed to provide the missing information.

We shall set the stage by making a connection between convection memory, or system memory in general, and vector autoregressive models with exogenous inputs (VARX). Description of this connection is mathematically straightforward but provides some intuition. We shall set the Kalman gain to zero as we intend to use the state-space models for simulation.

While the input and output data vectors are expressed using individual model layers as the bases, this need not be the case; any orthonormal transformations will not change the system identification problem described in section 3.

We define transformation matrices, $\mathbf{T_y}$ for outputs $\mathbf{y}$ and $\mathbf{T_f}$ for inputs $\mathbf{f}$ such that

$$\begin{aligned} \hat{\mathbf{y}}_t &= \mathbf{T_y}\mathbf{y}_t, \\ \hat{\mathbf{f}}_t &= \mathbf{T_f}\mathbf{f}_t, \quad \text{and} \\ \hat{\mathbf{e}}_t &= \mathbf{T_y}\mathbf{e}_t. \end{aligned} \tag{8}$$

We shall work in this transformed coordinate system. The state-space model, after setting the Kalman gain to zero, is now

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathbf{A}\mathbf{x}_t + \hat{\mathbf{B}}\hat{\mathbf{f}}_t \quad \text{and} \\ \hat{\mathbf{y}}_t &= \hat{\mathbf{C}}\mathbf{x}_t + \hat{\mathbf{e}}_t, \end{aligned} \tag{9}$$

where

$$\begin{aligned} \hat{\mathbf{C}} &\equiv \mathbf{T_y}\mathbf{C} \quad \text{and} \\ \hat{\mathbf{B}} &\equiv \mathbf{B}\mathbf{T}_\mathbf{f}^{\mathrm{T}} \end{aligned} \tag{10}$$

and superscript T indicates transpose. Define $\mathbf{z}_{t+1}$ as the convective sources of temperature and specific humidity over the time interval $[t\Delta, (t+1)\Delta]$:

$$\mathbf{z}_{t+1} \equiv \hat{\mathbf{y}}_{t+1} - \hat{\mathbf{y}}_t - \mathbf{T_y}\mathbf{T}_\mathbf{f}^{\mathrm{T}}\hat{\mathbf{f}}_t. \tag{11}$$

If $\mathbf{z}_{t+1}$ is determined by the current atmospheric state and forcing and an error term such that

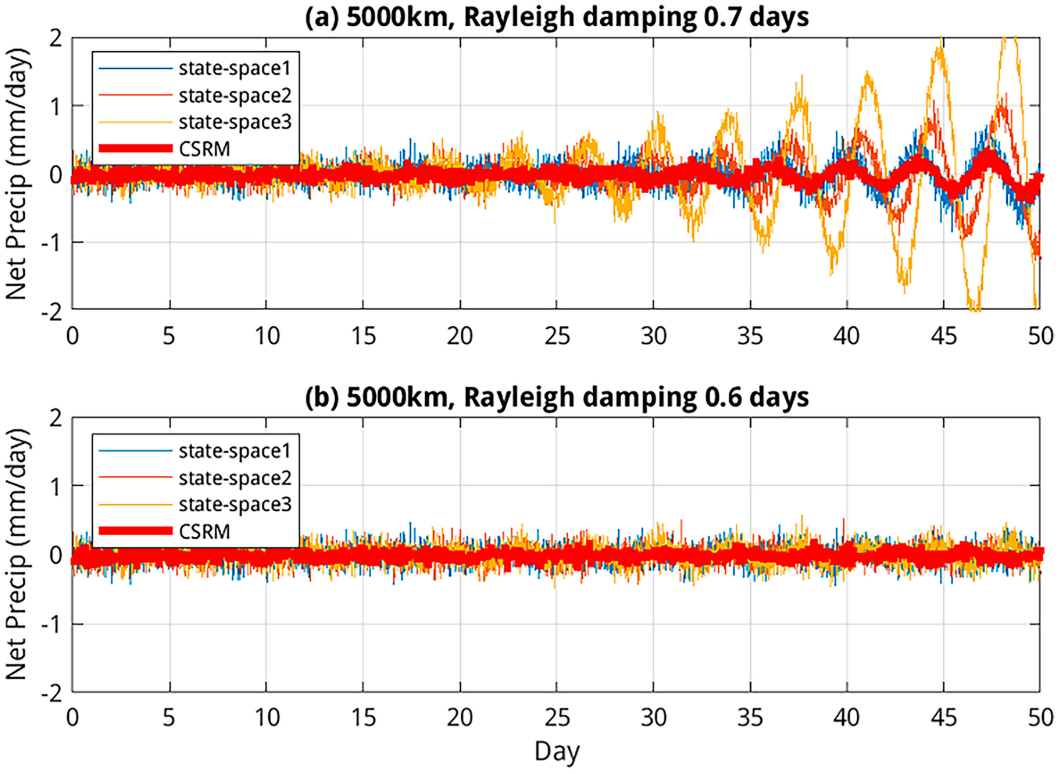

FIG. 6. As in Fig. 5, but for 2D gravity waves of 5000-km horizontal wavelength and Rayleigh damping of (a) 0.7 and (b) 0.6 days.

$$\mathbf{z}_{t+1} = (\boldsymbol{\Theta}_\mathbf{y} - \mathbf{I})\hat{\mathbf{y}}_t + (\boldsymbol{\Theta}_\mathbf{f} - \mathbf{T}_\mathbf{y}\mathbf{T}_\mathbf{f}^\mathrm{T})\hat{\mathbf{f}}_t + \hat{\mathbf{e}}_{t+1}, \tag{12}$$

we have a VARX1 (first-order VARX) process for $\hat{\mathbf{y}}$:

$$\hat{\mathbf{y}}_{t+1} = \boldsymbol{\Theta}_\mathbf{y}\hat{\mathbf{y}}_t + \boldsymbol{\Theta}_\mathbf{f}\hat{\mathbf{f}}_t + \hat{\mathbf{e}}_{t+1}, \tag{13}$$

where $\boldsymbol{\Theta}_\mathbf{y}$ may be viewed as a persistence matrix for $\hat{\mathbf{y}}$. If $\mathbf{z}_{t+1}$ is additionally affected by $\mathbf{z}_t$, that is, there is memory in convective tendencies, we have

$$\mathbf{z}_{t+1} = (\boldsymbol{\Theta}_\mathbf{y} - \mathbf{I})\hat{\mathbf{y}}_t + (\boldsymbol{\Theta}_\mathbf{f} - \mathbf{T}_\mathbf{y}\mathbf{T}_\mathbf{f}^\mathrm{T})\hat{\mathbf{f}}_t + \boldsymbol{\Theta}_\mathbf{z}\mathbf{z}_t + \hat{\mathbf{e}}_{t+1}, \tag{14}$$

where $\boldsymbol{\Theta}_\mathbf{z}$ is a persistence matrix for convective tendencies, representing, for example, inertia in convective updrafts and downdrafts. This leads to the following VARX2 (second-order VARX) process for $\hat{\mathbf{y}}$:

$$\hat{\mathbf{y}}_{t+1} = (\boldsymbol{\Theta}_\mathbf{y} + \boldsymbol{\Theta}_\mathbf{z})\hat{\mathbf{y}}_t + \boldsymbol{\Theta}_\mathbf{f}\hat{\mathbf{f}}_t - \boldsymbol{\Theta}_\mathbf{z}\hat{\mathbf{y}}_{t-1} - \boldsymbol{\Theta}_\mathbf{z}\mathbf{T}_\mathbf{y}\mathbf{T}_\mathbf{f}^\mathrm{T}\hat{\mathbf{f}}_{t-1} + \hat{\mathbf{e}}_{t+1}. \tag{15}$$

The above procedure can be continued. Define changes in the convective sources of temperature and humidity from time interval $[(t-1)\Delta, t\Delta]$ to time interval $[t\Delta, (t+1)\Delta]$:

$$\mathbf{a}_{t+1} \equiv \mathbf{z}_{t+1} - \mathbf{z}_t = \hat{\mathbf{y}}_{t+1} - 2\hat{\mathbf{y}}_t + \hat{\mathbf{y}}_{t-1} - \mathbf{T}_\mathbf{y}\mathbf{T}_\mathbf{f}^\mathrm{T}\hat{\mathbf{f}}_t + \mathbf{T}_\mathbf{y}\mathbf{T}_\mathbf{f}^\mathrm{T}\hat{\mathbf{f}}_{t-1}. \tag{16}$$

Equation (14) states that $\mathbf{a}_{t+1}$ can be determined by $\hat{\mathbf{y}}_t$, $\hat{\mathbf{f}}_t$, and $\mathbf{z}_t$ (plus the error term):

$$\mathbf{a}_{t+1} = (\boldsymbol{\Theta}_\mathbf{y} - \mathbf{I})\hat{\mathbf{y}}_t + (\boldsymbol{\Theta}_\mathbf{f} - \mathbf{T}_\mathbf{y}\mathbf{T}_\mathbf{f}^\mathrm{T})\hat{\mathbf{f}}_t + (\boldsymbol{\Theta}_z - \mathbf{I})\mathbf{z}_t + \hat{\mathbf{e}}_{t+1}. \tag{17}$$

If $\mathbf{a}_{t+1}$ is additionally affected by $\mathbf{a}_t$, that is, there is memory in changes in the convective sources of temperature and moisture, we have

$$\mathbf{a}_{t+1} = (\boldsymbol{\Theta}_\mathbf{y} - \mathbf{I})\hat{\mathbf{y}}_t + (\boldsymbol{\Theta}_\mathbf{f} - \mathbf{I})\hat{\mathbf{f}}_t + (\boldsymbol{\Theta}_\mathbf{z} - \mathbf{I})\mathbf{z}_t + \boldsymbol{\Theta}_\mathbf{a}\mathbf{a}_t + \hat{\mathbf{e}}_{t+1}, \tag{18}$$

where $\boldsymbol{\Theta}_\mathbf{a}$ is a persistence matrix for $\mathbf{a}$. Equation (18) gives a VARX3 (third-order VARX) process for $\hat{\mathbf{y}}$:

$$\begin{aligned}\hat{\mathbf{y}}_{t+1} = {} & (\boldsymbol{\Theta}_\mathbf{y} + \boldsymbol{\Theta}_\mathbf{z} + \boldsymbol{\Theta}_\mathbf{a})\hat{\mathbf{y}}_t - (\boldsymbol{\Theta}_\mathbf{z} + 2\boldsymbol{\Theta}_\mathbf{a})\hat{\mathbf{y}}_{t-1} + \boldsymbol{\Theta}_\mathbf{a}\hat{\mathbf{y}}_{t-2} \\ & + \boldsymbol{\Theta}_\mathbf{f}\hat{\mathbf{f}}_t - (\boldsymbol{\Theta}_\mathbf{z} + \boldsymbol{\Theta}_\mathbf{a})\mathbf{T}_\mathbf{y}\mathbf{T}_\mathbf{f}^\mathrm{T}\hat{\mathbf{f}}_{t-1} + \boldsymbol{\Theta}_\mathbf{a}\mathbf{T}_\mathbf{y}\mathbf{T}_\mathbf{f}^\mathrm{T}\hat{\mathbf{f}}_{t-2} + \hat{\mathbf{e}}_{t+1}.\end{aligned} \tag{19}$$

In theory, this procedure can be continued indefinitely, accounting for all possible memory effects. Truncating the procedure at a particular order is equivalent to assuming all higher-order differences of $\hat{\mathbf{y}}$ (with corresponding contributions from past $\hat{\mathbf{f}}$) are determined by (i.e., in statistical equilibrium with) current values of $\hat{\mathbf{y}}$ and its lower-order differences (again with corresponding contributions from past $\hat{\mathbf{f}}$).

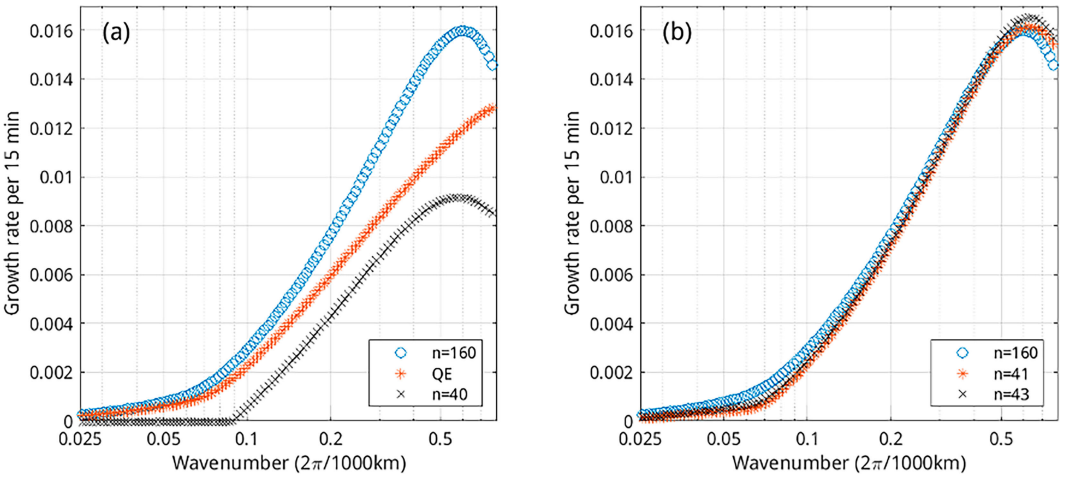

FIG. 7. (a) Growth rates of the system represented by Eq. (7) when the state-space models of orders 160 (blue circles) and 40 (black crosses) are used. The red asterisks are the growth rates when the state-space model in Eq. (7) is replaced by Eq. (21). (b) As in (a), but with state-space models of orders 41 (red asterisks), 43 (black crosses), and 160 (blue circles). Note that the size of input and output vectors is 40.

In Eqs. (15) and (19), coefficients of the $\hat{\mathbf{f}}_{t-1}$, $\hat{\mathbf{f}}_{t-2}$ terms and those of the $\hat{\mathbf{y}}_{t-1}$, $\hat{\mathbf{y}}_{t-2}$ terms are not linearly independent, because they only account for the passive role of $\hat{\mathbf{f}}$ in inferring convective sources based on budgets of the horizontal averages [see Eq. (11)]. However, there could be additional effects from forcing $\hat{\mathbf{f}}$ on convection beyond just the budgets, because, as we will see, horizontal averages do not fully describe the state of convection. Including additional terms to accommodate these, using Eq. (15) as an example, leads to a general VARX model:

$$\hat{\mathbf{y}}_{t+1} = (\mathbf{\Theta}_\mathbf{y} + \mathbf{\Theta}_\mathbf{z})\hat{\mathbf{y}}_t + \mathbf{\Theta}_\mathbf{f}\hat{\mathbf{f}}_t - \mathbf{\Theta}_\mathbf{z}\hat{\mathbf{y}}_{t-1} + [\mathbf{\Theta}_\mathbf{fl} - \mathbf{\Theta}_\mathbf{z}\mathbf{T}_\mathbf{y}\mathbf{T}_\mathbf{f}^\mathrm{T}]\hat{\mathbf{f}}_{t-1} + \hat{\mathbf{e}}_{t+1}. \quad (20)$$

We have designed a procedure to convert the identified state-space models to this VARX form that induces sparsity in the $\mathbf{\Theta}$ matrices to facilitate interpretation. While this procedure is key to the results that follow, it is technical in nature and its description is given in appendixes E–H. For the results to be presented, we use conversions to VARX3 models but have verified that, when the state-space models are converted to fourth-order VARX models, results do not change and the added terms are effectively zero. As explained in appendix H, we will focus on state-space models identified with CVA weighting and prediction focus.

a. *Models without memory*

We shall start with the case where model order n is 40, the size of the input and output vectors. In this case, we seek to maximize the sparsity of all $\mathbf{\Theta}$ matrices, including $\mathbf{\Theta}_\mathbf{y}$ and $\mathbf{\Theta}_\mathbf{f}$. Our procedure, described in appendixes E, F, G, and H, results in full rank $\mathbf{\Theta}_\mathbf{y}$ and $\mathbf{\Theta}_\mathbf{f}$, with all other $\mathbf{\Theta}$ matrices being zero. This means that outputs at time step $t + 1$ are computed with inputs and outputs at the time step t alone, i.e., there is no memory. We shall use $\mathbf{G}_{40}$ to denote the transfer function of this model, which represents the best model measured in L_2 norm without accounting for convection memory.

The approach of Kuang (2010) also has no memory because it can be written as

$$\mathbf{y}_{t+1} = (\mathbf{I} + \mathbf{M}^*)\mathbf{y}_t + \mathbf{f}_t \quad (21)$$

such that $\mathbf{y}_{t+1}$ only depends on $\mathbf{y}_t$ and $\mathbf{f}_t$. Here, $\mathbf{M}^*$ is the steady-state linear response function defined in Eq. (6) using parameters of the reference state-space model ($n = 160$) but with the sign of the positive eigenvalue flipped. This will be referred as the QE model. The transfer function in this case, $\mathbf{G}_\mathrm{QE}$, is

$$\mathbf{G}_\mathrm{QE}(\omega) = (\mathbf{I}e^{i\omega\Delta} - \mathbf{I} - \mathbf{M}^*)^{-1}. \quad (22)$$

In and only in this expression and Eq. (5), i denotes the imaginary unit. This model matches the state-space model at zero frequency but sacrifices accuracies at higher frequencies. The sign flip is inconsequential for the comparisons of the transfer functions shown in Fig. 8 but important for preserving the stability of the system. Because $\mathbf{M}^*$ describes how the system represented by Eq. (21) evolves with time, the system would be unstable if $\mathbf{M}^*$ had any positive eigenvalues. In contrast, the system represented by the reference state-space model is stable even though the matrix $\mathbf{M}$ derived from its parameters has a positive eigenvalue. This is because, in the state-space model, $\mathbf{M}$ only describes the system's steady-state response to forcing, not the system's time evolution.

Figures 8a and 8b show the errors in $\mathbf{G}_{40}$ relative to the transfer function of the $n = 160$ case, $\mathbf{G}_{160}$, used here as the "ground truth," which is justified given Fig. 2. Specifying the frequency ω (horizontal coordinate), the pressure level of the forcing (vertical coordinate) and whether it is a temperature or humidity forcing (left or right panel) selects a column of $\mathbf{G}_{40}(\omega) - \mathbf{G}_{160}(\omega)$, the difference matrix of the two transfer functions at that frequency. This column is a 40×1 vector that describes how responses of the temperature and humidity profiles to the specified forcing at

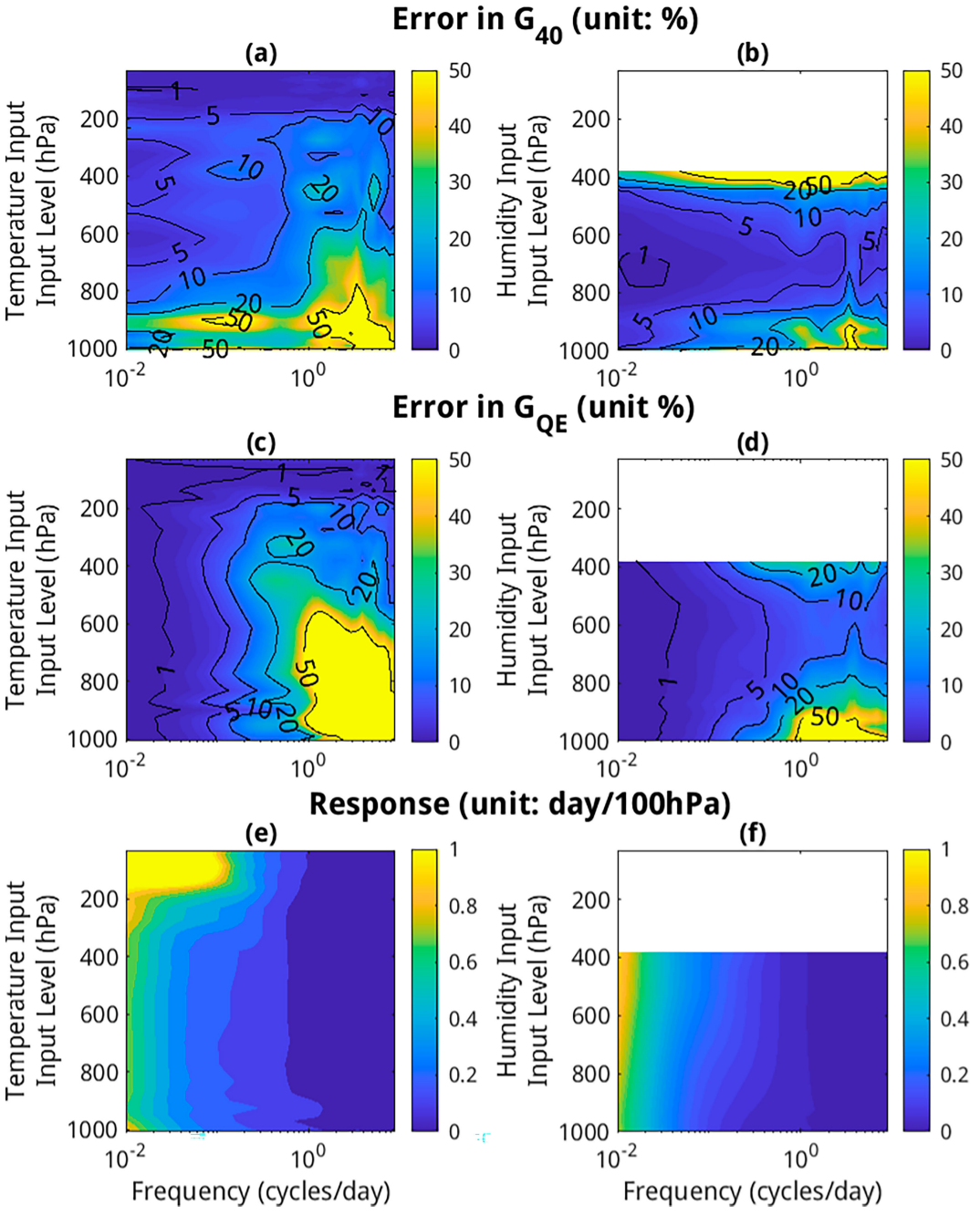

FIG. 8. Percentage errors in the transfer functions (a),(b) with model order 40 and (c),(d) with the QE assumption relative to the reference $n = 160$ LTI model for (left) temperature inputs and (right) humidity inputs. Also shown are amplitudes of the weighted L_2 norm of the responses to (e) temperature and (f) humidity inputs. The values in (e) and (f) are the weighted L_2 norm of the response in kelvins to a forcing with the amplitude of 1 K day^{-1} over a layer of 100 hPa in thickness at the specific pressure level (vertical axis) and frequency (horizontal axis).

the specified frequency differ between the two models. What is shown in Figs. 8a and 8b is the ratio, in percent, of the weighted L_2 norm of that column to the weighted L_2 norm of the corresponding column of $\mathbf{G}_{160}(\omega)$. The weighting is the same as that described in section 3c. Figures 8c and 8d show the parallel results for $\mathbf{G}_{\mathrm{QE}}$. The weighted L_2 norm of the different columns of $\mathbf{G}_{160}(\omega)$ for the range of frequencies considered are shown in Figs. 8e and 8f for reference.

In the case of $\mathbf{G}_{40}$, errors of 30%–50% or greater can occur at all frequencies. The steady-state matrix approach gives errors ~10% or less for periods 5 days or longer, which is consistent with the relatively good results obtained in previous

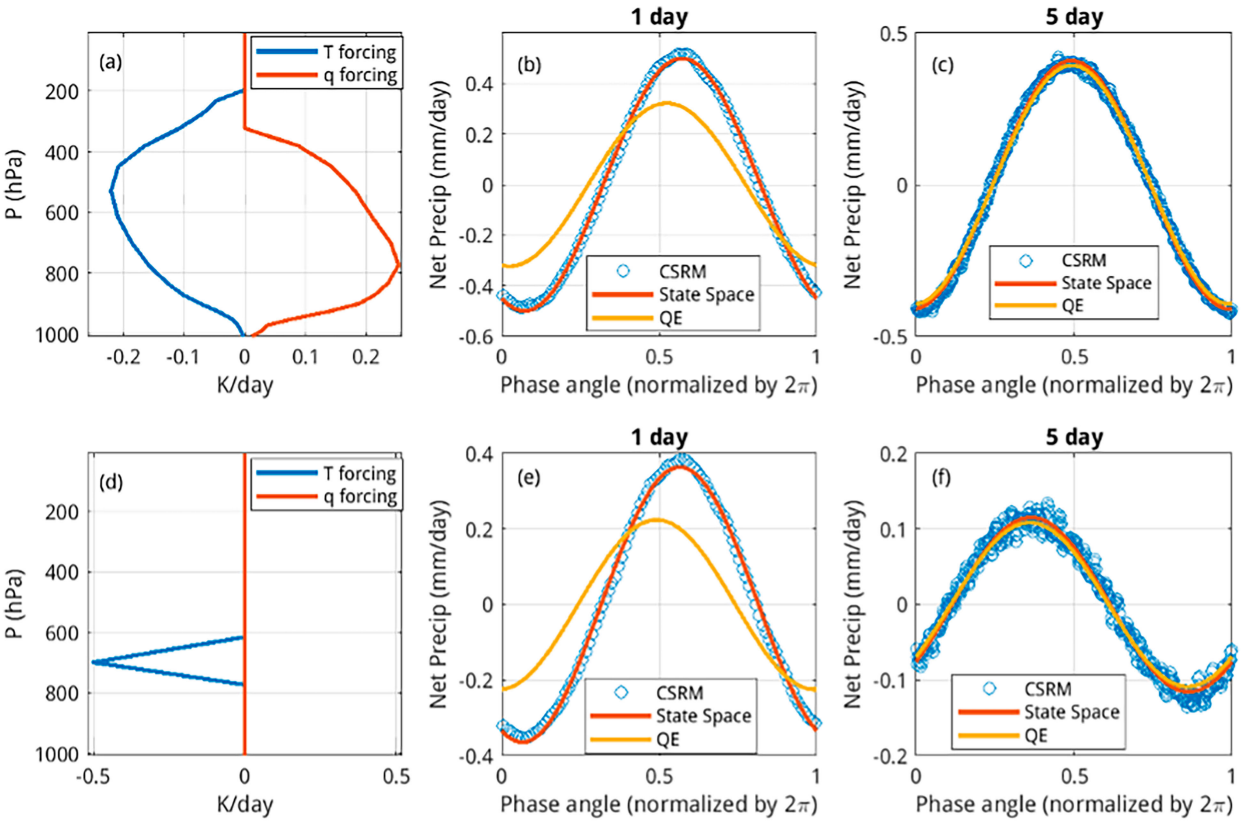

FIG. 9. (a) Structure of the sinusoidal forcings used, and responses in net precipitation when the forcing period is (b) 1 day and (c) 5 days. The phase of the forcing is such that the forcing at phase angle π is what is shown in (a) and the forcing at phase angle 0 is its negative. Results from the CSRM ensemble, the state-space model with $n = 160$, and the QE model [Eq. (21)] are shown in (b) and (c) in blue circles and red and yellow lines, respectively. (d)–(f) As in (a)–(c), but with the forcing structure shown in (d).

studies (e.g., Kuang 2010). Errors in $\mathbf{G}_{\mathrm{QE}}$ reach 20% or more at 2 days and can exceed 50% for periods of 1 day or shorter.

In Fig. 9, we further compare the net precipitation responses when forcing that is horizontally uniform and sinusoidal in time is applied to a 2048-member ensemble of the CSRM described in section 2, the reference state-space model, and the QE model of Eq. (21). The CSRM ensemble was run for 1200 days with the forcing and the last 1000 days were averaged to produce the net precipitation responses over one forcing period. For consistency, in all three cases (the CSRM ensemble, the state-space and QE models), net precipitation is computed from input and output based on conservation of water. In the upper panels, the applied forcing is a deep ascent/descent (Fig. 9a). This structure is similar to that used in Jones and Randall (2011) but with a much smaller amplitude. The response in net precipitation for a period of 5 days is closely in phase with the forcing, consistent with the notion of Arakawa and Schubert (1974) and well captured by the state-space model and the QE model. For a period of 1 day, the net precipitation response lags the forcing and substantial errors are seen in the QE model. These results are similar to those in Jones and Randall (2011).

For more general forcing structures, however, the response in net precipitation need not be in phase with the forcing even when the forcing period is long. Lower panels of Fig. 9 show that when a sinusoidal temperature forcing is applied at 700 hPa, the net precipitation response is not in phase with the forcing, but the more general notion of QE, as expressed in Eq. (21), still captures the response quite well. For a forcing period of 1 day, large errors are seen in the QE model. These results are not surprising given Fig. 8 but serve as more visual examples. Note that the forcing used here is stronger than that discussed in appendix A, which was fine-tuned to minimize nonlinearity, and some indication of nonlinearity can be seen in the CSRM results in Fig. 9. For example, in Figs. 9b and 9e, we observe that the positive phase of the net precipitation is slightly more peaked than the negative phase.

Figure 7a further compares the growth rates when the different models are coupled to 2D gravity waves. Results using the QE model [Eq. (21)] are close to those using the reference state-space model at lower horizontal wavenumbers but deviate from them substantially at higher horizontal wavenumbers. With the $n = 40$ state-space model, significant errors are seen across all horizontal wavenumbers. These results indicate that without accounting for convection memory and without prefiltering the data to emphasize lower frequency variations, as in the majority of current machine learning approaches to cumulus parameterization [with the notable exceptions of Han et al. (2020, 2023)], substantial errors likely exist in the convectively coupled waves.

TABLE 1. Nonzero eigenvalues of $\mathbf{\Theta_z}$ for a range of state-space model orders.

$n = 41$	$n = 42$	$n = 43$	$n = 44$	$n = 45$	$n = 46$	$n = 47$
0.85	0.83 ± 0.12i	0.89	0.92	0.96	0.95	0.96
		0.73 ± 0.22i	0.79 ± 0.30i	0.78 ± 0.21i	0.78 ± 0.21i	0.79 ± 0.23i
			0.51	0.47	0.50	0.70
				0.41	0.45	0.44
					0.08	0.41 ± 0.04i

b. Models with memory

Models with $n > 40$ include effects from past outputs and possibly past inputs, i.e., memory. As described in appendix H, we exclude $\mathbf{\Theta_y}$ and $\mathbf{\Theta_f}$ from the sparsity-inducing procedure, as we have found that they are always full rank. Even more specifically, we will focus on the interpretation of $\mathbf{\Theta_z}$, that is, memory in convective tendencies. Data-driven models such as those identified here in general combine different physical processes to achieve the smallest error given the model order. As an example, consider two physical processes that are correlated. Including a linear combination of the two in the model will achieve a smaller error than including only one of the two processes, which complicates the interpretation. If one physical process contributes much greater variance to the data than the other, then that process will dominate the linear combination. Our premise is that as the model order increases, processes added to the model will contribute progressively less variance to the data and processes that contribute most of the variance will eventually stabilize.

This appears to be the case as seen in Table 1, which lists the eigenvalues of $\mathbf{\Theta_z}$ when the model order increases from 41 to 47. While initially the leading eigenvalues (ranked by their modulus) vary considerably, for $n = 45$–47, the three leading eigenvalues have stabilized. As we further increase n, additional eigenvalues may stabilize, although we note that the reduction in the model error becomes more marginal (Fig. 1b). Furthermore, the computational cost of our procedure increases, as higher-order VARX models would be needed, and the sparsity-inducing procedure also becomes less effective. We limit ourselves in this paper to the first three eigenmodes and leave the interpretation of additional eigenmodes of $\mathbf{\Theta_z}$ as well the eigenmodes of the other $\mathbf{\Theta}$ matrices to future work.

The leading eigenvalue is ~0.96. The associated eigenvector for $n = 45$, 46, 47 shows reasonable consistency (Fig. 10). We interpret this mode as the deep convective mode and the eigenvalue indicates that this mode of convective tendency will decay in amplitude by ~4% every 15 min; that is, it has an e-folding time of ~5 h. The second and third eigenvalues form a conjugate pair with an e-folding time of ~1 h and a period of 6 h. Figure 11 visualizes how convective tendencies associated with the second and third eigenvectors evolve with time with the exponential decay suppressed in the graphic to make phase structure clearer. The time evolution resembles the congestus to deep convection to stratiform life cycle emphasized in, for example, Mapes (2000). The 6-h period also offers a plausible explanation of the peaks at 6 h in the stochastic noise spectra seen in Fig. 2. The finding that persistence and life cycle of the convective elements contribute to convection memory and the character of the stochastic noise is not surprising. However, the data-driven identification and the level of quantification that can be obtained with our approach is a new addition to the studies of cumulus convection.

It is left to future study to develop interpretations of additional elements of convection memory, but the picture painted in Figs. 10 and 11 appears to capture a majority of the memory effect. Figure 7b shows that including memory from only a few modes improves the growth rates of the convectively coupled waves substantially.

6. Conclusions and further discussion

In this paper, we take a time series analysis perspective and use methods in system identification to obtain LTI state-space models that describe how horizontal averages of temperature and humidity of a large cumulus ensemble evolve with time under small forcing. The identified state-space models, with sufficient model order, accurately represent the behavior of the cumulus ensemble in terms of its transfer function, the noise model, and its behavior when coupled with 2D gravity waves. The highest accuracy is obtained when the model order is greater than about 150. These state-space models extend the steady-state linear response function approach used in past studies such as Kuang (2010, 2018) by removing the assumption that the cumulus ensemble is in statistical equilibrium with its large-scale environment.

A novel procedure is then developed to convert the state-space models into a sparse VARX form to facilitate physical interpretation. The procedure is applied to the identified state-space models with a range of model orders. State-space models that, when converted to the sparse VARX form, use only the current input and output $\mathbf{y}_t$ and $\mathbf{f}_t$ have transfer functions that contain substantial (30%–50% or more) errors for all frequencies. The QE approach, another memoryless model, works well at lower frequencies but becomes inaccurate at higher frequencies. Our results also indicate that without accounting for convection memory and without prefiltering the data to emphasize lower frequency variations, data-driven cumulus parameterizations likely contain substantial errors in the convectively coupled waves. Last, a robust and sparse representation of the leading convection memory effects is found and resembles the deep convective heating mode and the congestus to deep convection to stratiform convective life cycle previously emphasized in the literature. The deep convective heating mode is found to have an e-folding

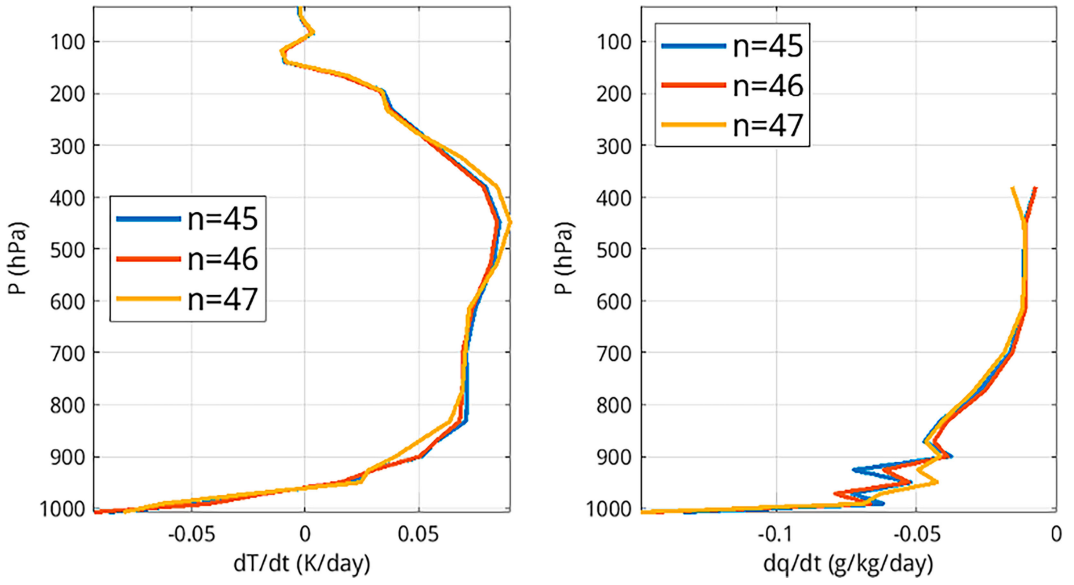

FIG. 10. Convective (left) temperature and (right) humidity tendency structures associated with the leading eigenvector (with the largest eigenvalue in modulus) of $\mathbf{\Theta_z}$ using the state-space model with model orders of $n = 45$ (blue), 46 (red), and 47 (yellow).

time of ~5 h and the congestus to deep convection to stratiform oscillation is found to have a period of ~6 h with an *e*-folding time of 1 h.

An immediate limitation of the present effort, aside from the limitations of the CSRM used to perform the experiments in terms of model physics, setup, and numerics, is that it is limited to small perturbations around a reference mean state. While the results here are already relevant to phenomena such as the convectively coupled waves, extension to the nonlinear regime is needed to address a wider range of problems,

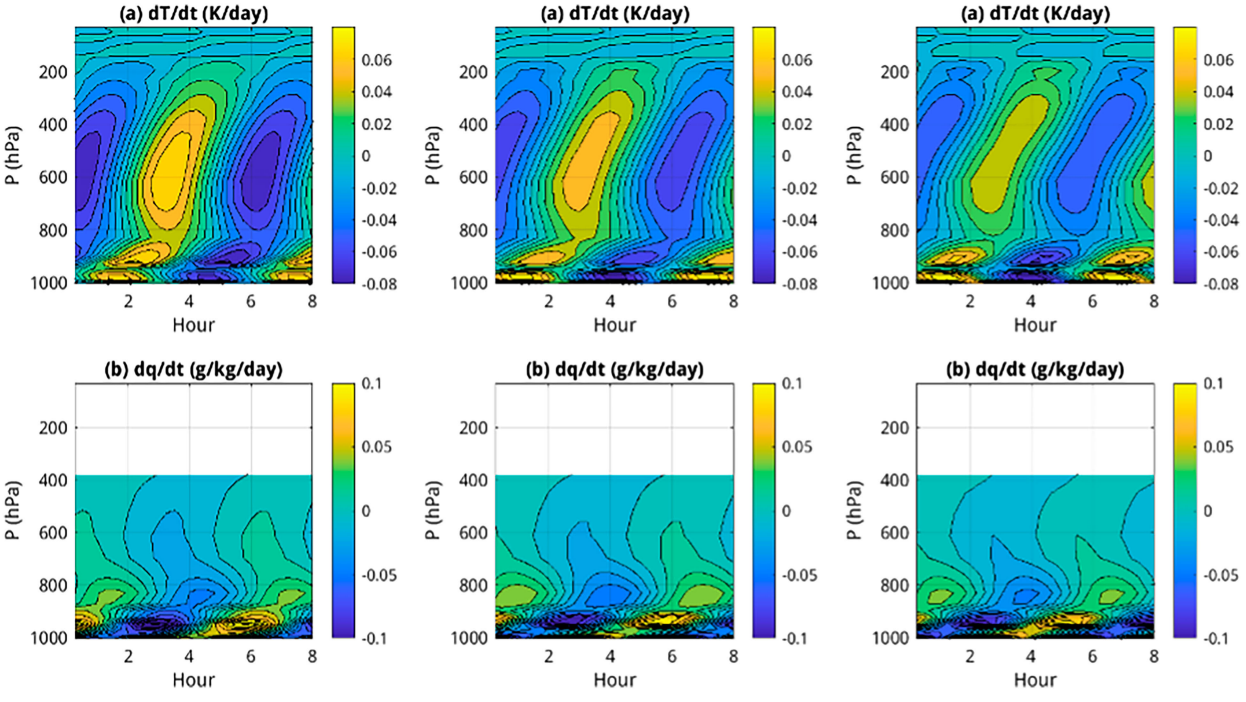

FIG. 11. Time evolutions of convective (a) temperature and (b) humidity tendency structures associated with the second and third eigenvectors of $\mathbf{\Theta_z}$ using the state-space model with model orders of $n =$ (left) 45, (middle) 46, and (right) 47. The two eigenvectors form a complex conjugate pair, and the decay of the eigenmodes has been removed in the figure to emphasize the oscillatory aspect.

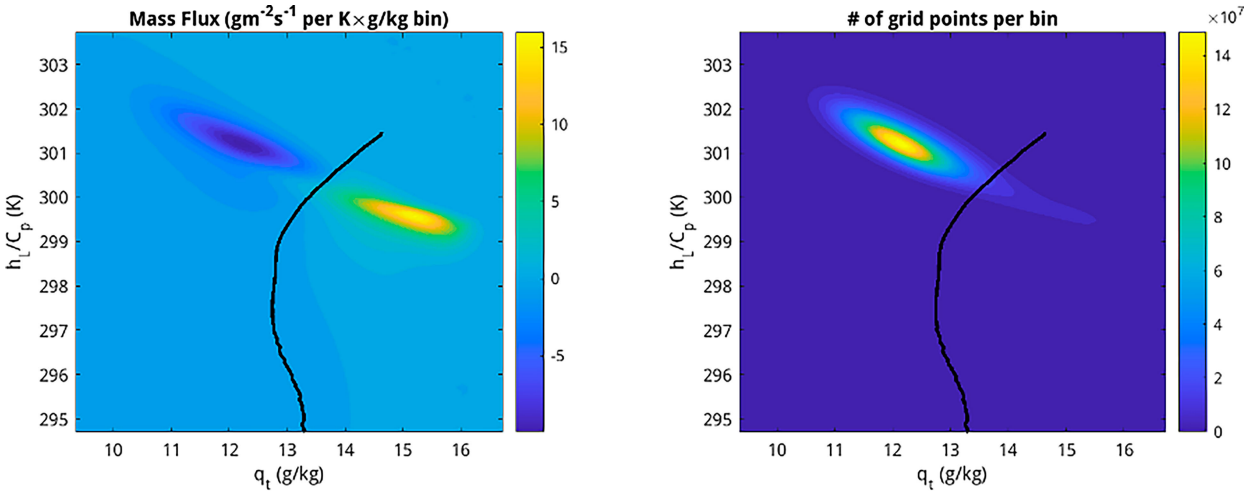

FIG. 12. (left) Mass flux as a function of liquid water static energy divided by the specific heat h_L/c_p and total nonprecipitating water q_t. (right) Joint histogram of h_L/c_p and q_t. The thick line demarcates the boundary between cloudy (average cloud water in the bin greater than 10^{-5} g kg^{-1}) and noncloudy bins. The bin size is ~0.09 K for h_L/c_p and ~0.1 g kg^{-1} for q_t.

including the parameterization of cumulus convection. One avenue that we are pursuing is to use the identified state-space models to initialize a recurrent neural network, which has a similar construct as a state-space model but can provide nonlinear representations (with the trade-off of having more complex and less well understood model estimation), and gradually ramp up the forcing amplitude in our experiments to extend to the more nonlinear regimes. Results from such efforts will be reported in the future.

In multivariate time series analysis, model equivalency and nonuniqueness are well known. Existing approaches to select a unique model among many equivalent models, known as model specification, include the scalar component model approach and the echelon form (e.g., Tiao and Tsay 1989; Athanasopoulos et al. 2012). Our procedure of identifying state-space models from the time series and then converting the state-space models to VARX models, as described in the appendixes, can be easily extended to VARMAX models and is a new addition to such efforts. Much uncertainty, however, remains. The assumption that greater sparsity provides more interpretability, while reasonable, remains an assumption; the sparsity-inducing procedure remains approximate; confidence in the interpretations given in section 5b also partially relies on their consistency with our prior knowledge of convective life cycle. More work is needed to improve the discovery of new physics in purely data-driven models.

As noted at the beginning of section 5, the equations solved by the CSRM are autonomous. Memory only arises because we have reduced the prognostic variables of the CSRM to horizontally averaged profiles of temperature and humidity—our **y**. This reduction loses information needed to determine how the

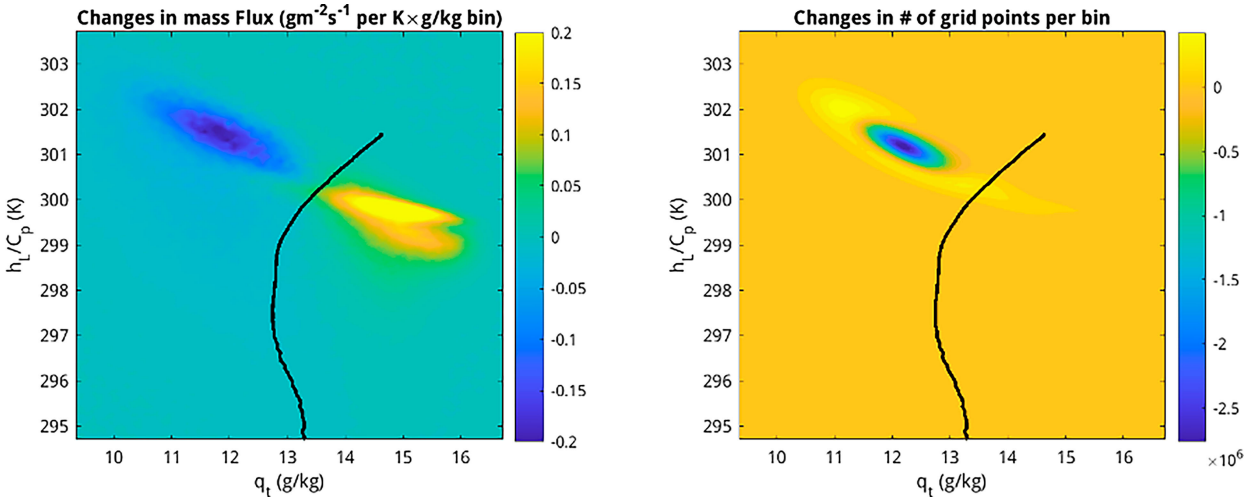

FIG. 13. As in Fig. 12, but for changes associated with a steady-state forcing that aligns with the polished eigenvector of **M** with the positive eigenvalue, as shown in Fig. 4.

system evolves with time, and past values of **y** and **f** are used to provide the missing information. In this sense, past values of **y** and **f** serve as proxies of the missing information stored in the state vector **x**.

An alternative approach is to extend the output vector **y** to include additional statistics of the cumulus ensemble needed to determine its evolution. One candidate potentially useful to include is aspects of the joint probability distribution functions (PDFs). As an initial illustration of this potential, we show here that the PDFs or histograms can provide a more physical picture for the seemingly counter intuitive results of a positive eigenvalue of the steady-state linear response function, as described in section 4d.

To start, we show in Fig. 12 the reference joint histogram of liquid water static energy, h_L (divided by c_p so it is in temperature unit) and total nonprecipitating water, q_t, as well as the vertical mass flux for each of the bins, at the cloud base (~900 hPa). As expected, the upward mass fluxes are carried by a small fraction of the grid points that are at the moister and colder end of the distribution and are cloudy. The steady-state response to forcing that aligns with the eigenvector of **M** with the positive eigenvalue is shown in Fig. 13. In response to horizontally uniform warming and moistening in the boundary layer and cooling and moistening in the free troposphere, upward and downward mass fluxes increase, so does the number of cloudy grid points at the moister and cooler end of the joint distribution, consistent with the stronger convective overturning needed to balance the forcing. However, there are also more grid points at the drier and warmer end of the joint distribution, leading the horizontal average to be drier and warmer, as seen in Fig. 4.

Figure 14 further shows the histogram of h_L at the lowest model layer and its response to the aforementioned forcing. The reference histogram shows a long low-temperature tail due to cold pools. A horizontally uniform warming tendency shifts the histogram toward higher h_L without changing its shape. The steady-state response, however, shows a wider distribution with more grid points at both the high- and low-temperature ends. While the horizontal average is colder in response to a warming tendency near the surface (due to more extensive cold pools associated with stronger convection), there are also more parcels at the high-temperature end, allowing for more near-surface air parcels to participate in deep convection. The results shown in Fig. 4 can thus be interpreted physically with the aid of auxiliary information from the joint histograms.

The need to include variables in addition to horizontal averages has been recognized in the past. For example, Mapes and Neale (2011) proposed adding a prognostic variable to represent convection organization. More recent work using machine learning to parameterize convection found that including some measure of convection organization improves the prediction of precipitation extremes (Shamekh et al. 2023). The identified state-space models already augment **y** through its use of state vector to represent the latent space, but in a "black box" manner. With the sparse VARX form of the identified state-space models, we can isolate the contributions needed from variables in addition to current **y** and **f** in a more systematic way and select components from the joint PDF and other statistics accordingly. Such efforts could help us identify and quantify, for examples, the

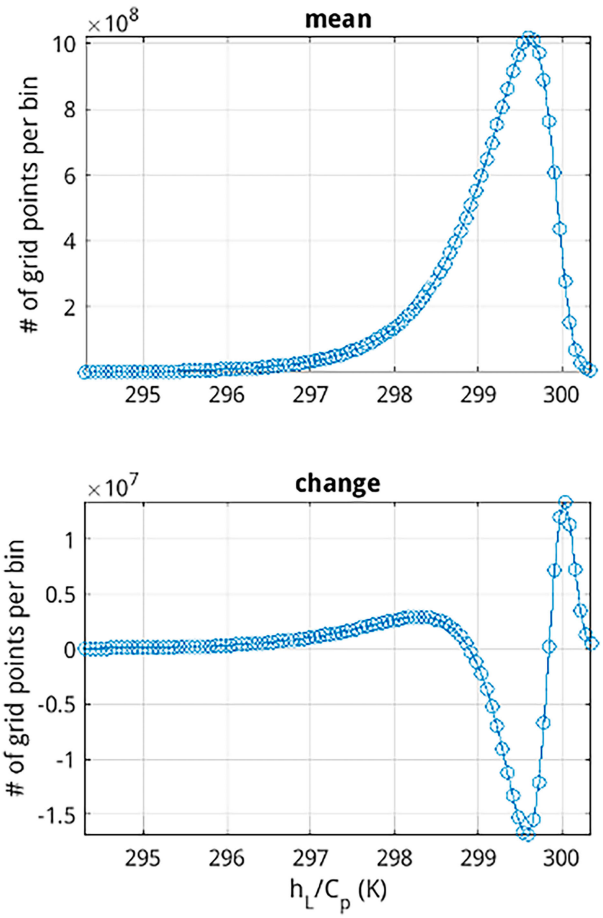

FIG. 14. (top) The reference mean histogram of h_L/c_p at the lowest model layer. The bin size is ~0.06 K. (bottom) Changes in the histogram associated with a steady-state forcing that aligns with the polished eigenvector of **M** with the positive eigenvalue, as shown in Fig. 4.

aspects of the cumulus ensemble that are responsible for the persistence in its deep convective heating and for its evolution from congestus to deep convective to stratiform heating, without contributions from the horizontal averages. Further connecting with cumulus parameterizations with assumed PDFs (e.g., Golaz et al. 2002a,b) could also be a valuable direction to explore.

Acknowledgments. The author thanks Brian Mapes and an anonymous reviewer for their careful reviews and Marat Khairoutdinov for making the SAM model available. This work was supported by NSF Grant AGS-1759255 and by the Office of Biological and Environmental Research of the U.S. DOE under Grant DE-SC0022887 as part of the Atmospheric System R program. The Harvard Odyssey cluster provided the computing resources for this work.

Data availability statement. All simulations were performed with the System for Atmospheric Modeling, developed and maintained by Dr. Marat Khairoutdinov at Stony Brook University and available online (http://rossby.msrc.sunysb.edu/~marat/

SAM.html). Other software used is from the MATLAB System Identification toolbox. Time series used for training and validation, as well as the identified models, are available through Harvard Dataverse (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/IY3CJJ).

APPENDIX A

Generation of the Input Sequences

To generate the 1000-day input sequences, we first generate 200-day time sequences using randomly phased multisines, and then repeat them five times. More specifically, the time sequence for the ith component of $\mathbf{f}$ is given by

$$f_t^{(i)} = \frac{2}{\sqrt{N}} \sum_{l=1}^{N/2} F_l^{(i)} \cos\left[\frac{2\pi l t}{N} + \varphi_l^{(i)}\right]$$
$$t = 1, 2, \ldots, 5N - 1, 5N, \tag{A1}$$

where $N = 200$ days/Δ and $\varphi_l^{(i)}$ is the phase shift of the lth Fourier component of the ith component of $\mathbf{f}$. The different phase shifts are independently drawn from a uniform distribution over $[0, 2\pi]$. With $\Delta = 15$ min, N is sufficiently large such that $f_t^{(i)}$ approaches Gaussian noise (within each 200-day period), and the time series of $f_t^{(i)}$ and $f_t^{(j)}$ are uncorrelated when $i \neq j$. The choice to make forcing in temperature and humidity and in different layers uncorrelated is made for simplicity. Having an optimized nondiagonal covariance matrix for $\mathbf{f}$ could potentially produce higher SNRs and smaller nonlinear distortions but the optimization is highly nontrivial and is not pursued here.

The forcing amplitude $F_l^{(i)}$ is chosen to balance nonlinear distortion and SNR. We first made a 1000-day run using forcing time sequences given by Eq. (A1) with $F_l^{(i)} = F^{(i)}$ for odd l and $F_l^{(i)} = 0$ for even l. $F^{(i)}$ is set to 0.02 K/15 minutes if the ith component of $\mathbf{y}$ represents the temperature of a layer. If the ith component of $\mathbf{y}$ represents the specific humidity of a layer, $F^{(i)}$ is set to the reference specific humidity of that layer divided by 10 days. The relative amplitudes of temperature and humidity forcing at different layers were chosen based on past experiences.

Discarding the first 200-day period, the remaining four 200-day periods, having the same forcing and unaffected by transients (initial adjustments), are averaged. The stochastic noise is estimated by the deviation of individual periods from the average. Since forcing is zero at even frequencies, greater power at these frequencies than that of stochastic noise at nearby odd frequencies indicates the presence of nonlinear distortion, assuming that nearby frequencies behave similarly. With the aforementioned forcing amplitudes, nonlinear distortion is readily detected above the level of stochastic noise (Fig. A1a). The ability to quantify the nonlinear distortion is a major motivation for using the randomly phased multisines (Schoukens et al. 2016).

Assuming that the nonlinear distortion is dominated by quadratic terms, estimates can be made to adjust the forcing amplitude such that nonlinear distortion does not exceed the stochastic noise. Based on such estimates, we set

$$F_l^{(i)} = F^{(i)} \Big/ \left\{3.5 - 1.25 \max\left[\log_{10}\left(\frac{3l}{400}\right), 0\right]\right\}$$
$$l = 1, 3, 5, \ldots, \frac{N}{2} - 1. \tag{A2}$$

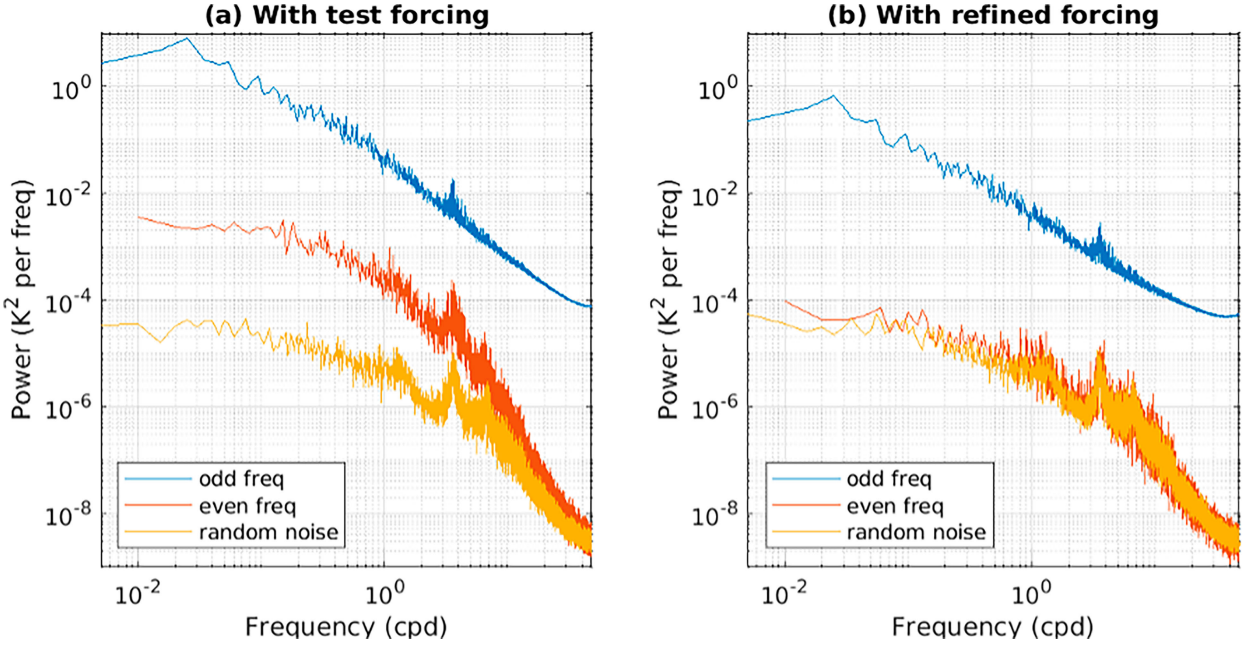

FIG. A1. (a) Mass-weighted average of the power spectra of the output signal $\mathbf{y}$ at odd (blue) and even frequencies (red) and of random noise (yellow) when the CSRM ensemble is forced with the test odd-frequency-only forcing, as described in appendix A. (b) As in (a), but with refined forcing amplitude described in Eq. (A2).

Nonlinear distortion is no longer detected in simulations with these input amplitudes (Fig. A1b). In the simulations that we use to identify the state-space model, we divide the forcing described in Eq. (A2) by $\sqrt{2}$ and apply it to all l values from 1 to $N/2$ such that the forcing contains the same total spectra power. This uses twice as many frequencies and reduces the stochastic nonlinear distortion and is preferable to Eq. (A2).

APPENDIX B

Hyperparameter Search

The subspace method contains several user (or hyperparameter) choices. First, one can choose which weighting matrices to apply to estimates of the extended observability matrix in the algorithm. We will focus on the following two choices: multivariate output-error state space (MOESP) (Verhaegen 1994) and CVA (Larimore 1990). Because the CVA weighting accounts for the error covariance and MOESP does not, we have found that, for parameter estimation, the CVA weighting works well with prediction error minimization while the MOESP weighting works well with simulations error minimization, and the other pairings produce significantly worse results. Another user choice is the numbers of past and future inputs and outputs to use in the estimation algorithm. Using a sufficiently large number of past inputs and outputs is important for capturing the relevant information, and using a sufficiently large number of future inputs and outputs is important for constraining long time scales of the evolution matrix $\mathbf{A}$ and avoiding stability issues. On the other hand, making the number of past and future inputs and outputs too large will add noise and computational expense. Results for a few choices are presented in this appendix.

The black symbols in Fig. 1a give a broad view of how the percentage validation error varies with model order using the CVA weighting with prediction focus and with the number of past and future inputs and outputs used chosen automatically by MATLAB based on the Akaike information criterion (AIC). That the validation error increases as the model order approaches 300 is due to MATLAB using too few past inputs and outputs (<5) in the parameter estimation for these model orders. Using a greater number of past inputs and outputs (green symbols in Fig. 1c) removes this error increase. Therefore, it appears that the error stabilizes beyond model order of ~150.

Figure 1b is a close-up on the small model order portion. MATLAB's automatic selection uses a smaller number of past inputs and outputs as the model order becomes small. To test the sensitivity to this, we reestimated the parameters for model orders less than 48 using 96 past and 72 future inputs and outputs, values chosen based on the AIC for $n = 48$, and the results are shown in red. The errors are largely similar except a substantial reduction at $n = 43$. Using even larger numbers of past and future inputs and outputs had little effects on the errors. The cases shown in red are those used in section 5 for physical interpretation of convection memory.

We have also included the results from MOESP weighting with simulation focus in Fig. 1. Their errors are smaller than those from CVA weighting of the same model order when the model order is low and converge to the CVA results when model order reaches ~100. For physical interpretation described in section 5, we use the models with CVA weighting for reasons described in appendix H.

APPENDIX C

Formulas for Lag Noise Covariance Matrices

Without forcing, variations in the system are entirely due to stochastic noise:

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{K}\mathbf{e}_t \quad \text{and}$$
$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{e}_t. \tag{C1}$$

The lag-0 covariance in $\mathbf{x}$, $\mathbf{S}_\mathbf{x}$, in this case is given by the discrete Lyapunov equation:

$$\mathbf{S}_\mathbf{x} = \mathbf{A}\mathbf{S}_\mathbf{x}\mathbf{A}^\mathrm{T} + \mathbf{K}\mathbf{R}\mathbf{K}^\mathrm{T}, \tag{C2}$$

where $\mathbf{R}$ is again the covariance matrix of $\mathbf{e}$. The lag-0 covariance in $\mathbf{y}$, $\mathbf{S}_\mathbf{y}$, is

$$\mathbf{S}_\mathbf{y} = \mathbf{C}\mathbf{S}_\mathbf{x}\mathbf{C}^\mathrm{T} + \mathbf{R}. \tag{C3}$$

Similar derivations give the lag-j covariance in $\mathbf{y}$ as

$$\mathbf{S}_\mathbf{y}^{(j)} = \mathbf{C}\mathbf{A}^j\mathbf{S}_\mathbf{x}\mathbf{C}^\mathrm{T} + \mathbf{C}\mathbf{A}^{j-1}\mathbf{K}\mathbf{R}. \tag{C4}$$

APPENDIX D

Refinement of the Eigenvector of the Steady-State Linear Response Matrix Associated with the Positive Eigenvalue

The amplitudes of the positive eigenvalue and the smallest (in absolute value) negative eigenvalue of the steady-state linear response function matrix $\mathbf{M}$ differ by a factor of ~440. When we force the CSRM ensemble with the eigenvector associated with the positive eigenvalue, errors in that eigenvector can potentially be amplified by as much. Furthermore, the state-space model excluded specific humidity above ~350 hPa from the state vector for simplicity. This is adequate for most purposes but may not suffice here. For this reason, we have further refined this eigenvector through the following procedure, making use of the fact that there is only one positive eigenvalue.

In this appendix only, the input and output vectors are extended to include T and q of all except the top two model levels. If the forcing vector $\mathbf{f}$ is the eigenvector associated with the positive eigenvalue, the sum of the normalized steady-state output vector $\mathbf{y}$ and the normalized $\mathbf{f}$ should be zero. A nonzero sum indicates deviations of the forcing vector from the true eigenvector. We therefore adjust the direction of the forcing vector using the following equation:

$$\mathbf{f}_{t+1} = \frac{\mathbf{f}_t}{\|\mathbf{f}_t\|} - \frac{\Delta}{\tau}\left(\frac{\mathbf{f}_t}{\|\mathbf{f}_t\|} + \frac{\mathbf{y}_t}{\|\mathbf{y}_t\|}\right), \tag{D1}$$

where Δ is 15 min and τ is set to 200 days to ensure that $\mathbf{y}$ is close to the steady-state response (the adjustment to equilibrium takes about 20 days) and that adjustment in $\mathbf{f}$ is small to avoid artificial amplification. The forcing $\mathbf{f}$ is initialized with the eigenvector of $\mathbf{M}$ with the positive eigenvalue, padded with zero for humidity values above 350 hPa. This polished eigenvector is used in the lower panels of Fig. 4.

APPENDIX E

Overview of the Conversion of State-Space Models to an Interpretable Form

We start with the approach by Phan et al. (1998) to convert a state-space model to a VARX model with p past inputs and outputs. We shall neglect the Kalman gain term in Eq. (2), as we intend to use the model for simulation instead of prediction. Phan et al.'s framework can be extended to include the Kalman gain term to convert a state-space model into a VARMAX (vector autoregressive moving average with external input) model. We will not present that result here and will focus on the VARX formulation.

First, define the controllability matrix $\mathscr{C}_p$, observability matrix $\mathscr{O}_p$, and Toeplitz matrix $\mathscr{T}_p$:

$$\mathscr{C}_p \equiv [\mathbf{A}^{p-1}\hat{\mathbf{B}}, \quad \ldots, \quad \mathbf{A}\hat{\mathbf{B}}, \hat{\mathbf{B}}],$$

$$\mathscr{O}_p \equiv \begin{bmatrix} \hat{\mathbf{C}} \\ \hat{\mathbf{C}}\mathbf{A} \\ \vdots \\ \hat{\mathbf{C}}\mathbf{A}^{p-1} \end{bmatrix}, \quad \text{and}$$

$$\mathscr{T}_p \equiv \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \hat{\mathbf{C}}\hat{\mathbf{B}} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \hat{\mathbf{C}}\mathbf{A}\hat{\mathbf{B}} & \hat{\mathbf{C}}\hat{\mathbf{B}} & \mathbf{0} & \ddots & \mathbf{0} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \hat{\mathbf{C}}\mathbf{A}^{p-2}\hat{\mathbf{B}} & \cdots & \hat{\mathbf{C}}\mathbf{A}\hat{\mathbf{B}} & \hat{\mathbf{C}}\hat{\mathbf{B}} & \mathbf{0} \end{bmatrix}. \tag{E1}$$

Following Phan et al. (1998), with a matrix $\mathbf{N}$ that satisfies[E1]

$$\hat{\mathbf{C}}\mathbf{A}^p + \mathbf{N}\mathscr{O}_p = 0, \tag{E2}$$

the state-space model [Eq. (9)] can be converted to a one-step-ahead input–output model:

$$\hat{\mathbf{y}}_t = \sum_{i=1}^{p} \Phi_i \hat{\mathbf{y}}_{t-i} + \sum_{i=1}^{p} \Psi_i \hat{\mathbf{f}}_{t-i} + \hat{\mathbf{e}}_t, \tag{E3}$$

with

$$[\Phi_p \quad \ldots \quad \Phi_2 \quad \Phi_1] = -\mathbf{N} \quad \text{and}$$

$$[\Psi_p \quad \ldots \quad \Psi_2 \quad \Psi_1] = \mathbf{C}\mathscr{C}_p + \mathbf{N}\mathscr{T}_p. \tag{E4}$$

[E1] Equation (E2) here is Eq. (5) of Phan et al. (1998) multiplied by $\hat{\mathbf{C}}$. This change slightly simplifies our discussion.

The steps so far are the same as those in Phan et al. (1998) except that there is no direct feedthrough. Including a direct feedthrough term poses no difficulty for the discussions below.

When $n < mp$, Eq. (E2) is underdetermined, so $\mathbf{N}$ is nonunique. The problem then is how to choose $\mathbf{N}$ to simplify the interpretation. This is where we deviate from Phan et al. (1998), which uses the pseudoinverse to minimize the Frobenius norm of $\mathbf{N}$. We instead minimize the sparsity inducing L_1 norm of the vector that contains all the elements of the persistent matrices ($\mathbf{\Theta_y}$, $\mathbf{\Theta_z}$, $\mathbf{\Theta_a}$, etc.), as well as $\mathbf{\Theta_f}$ and the matrices describing additional contributions from $\mathbf{f}_{t-1}$, $\mathbf{f}_{t-2}$, and so on (see section 5 for definitions of these matrices). While orthonormal transformations of $\mathbf{y}$ and $\mathbf{f}$ do not change the Frobenius norms of the coefficient matrices, they do change the L_1 norm. Therefore, the minimization problem is to find $\mathbf{T_y}$, $\mathbf{T_f}$ and $\mathbf{N}$ given by

$$\underset{\mathbf{N},\mathbf{T_y},\mathbf{T_f}}{\operatorname{argmin}}\{\|\mathrm{col}([\mathbf{\Theta_y} \quad \mathbf{\Theta_z} \quad \mathbf{\Theta_a} \quad \ldots])\|_1 + \|\mathrm{col}([\mathbf{\Theta_f} \quad \mathbf{\Theta_{f1}} \quad \mathbf{\Theta_{f2}} \quad \ldots])\|_1\}, \tag{E5}$$

where col$(\cdot)$ is an operator that stacks the columns of a matrix on top of each other to form a column vector and $\|\cdot\|_1$ denotes the L_1 vector norm. To truly minimize the number of nonzero parameters in the model, which is assumed to allow for greater interpretability, one would need to minimize the L_0 norm. However, that problem is prohibitively expensive given the dimensions of the present system and minimizing L_1 norm is used as a proxy. Advances that led to compressed sensing have shown that solutions to L_1-norm minimization also minimize the L_0 norm for large random matrices under certain conditions (e.g., Candès et al. 2006; Donoho 2006). However, those conditions are not met in our case. Earlier results that rely on coherence (see review in, e.g., Donoho 2006) are also too restrictive to be applied here. Therefore, we do not have a mathematical proof that the solution minimizing the L_1 norm is the unique optimally sparsest solution. Nevertheless, the results appear sufficiently sparse to allow for physical interpretation.

The relative weights between the terms in Eq. (E5) are arbitrary. We have found that the two terms are of similar magnitude and the results are insensitive to the relative weights so have simply used weights of 1, as indicated in Eq. (E5).

This minimization problem is solved in two steps. First, given the orthogonal transformation matrices $\mathbf{T_y}$ and $\mathbf{T_f}$, the minimization problem is formulated as a linear programming problem (see appendix F for the formulation) and solved using the simplex algorithm, as implemented in MATLAB. The minimization over $\mathbf{T_y}$ and $\mathbf{T_f}$ is nonlinear and is done using the quasi-Newton method, facilitated by the fact that the gradient can be computed based on the Lagrange multipliers from the linear programming step. The details are given in appendix G.

APPENDIX F

Formulation of the Linear Programming Problem

For this appendix, our starting point is Eq. (9) with $\mathbf{T_y}$ and $\mathbf{T_f}$ given. The following describes the case with $p = 2$. Cases

with larger p values can be dealt with similarly. For $p = 2$, the controllability, observability, and Toeplitz matrices are

$$\mathscr{C}_2 \equiv [\,\mathbf{A}\hat{\mathbf{B}} \quad \hat{\mathbf{B}}\,],$$
$$\mathscr{O}_2 \equiv \begin{bmatrix} \hat{\mathbf{C}} \\ \hat{\mathbf{C}}\mathbf{A} \end{bmatrix}, \quad \text{and}$$
$$\mathscr{T}_2 \equiv \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \hat{\mathbf{C}}\hat{\mathbf{B}} & \mathbf{0} \end{bmatrix}. \tag{F1}$$

From Eqs. (E2) and (E4), which follow Phan et al. (1998), we have

$$\mathbf{\Psi}_1 = \hat{\mathbf{C}}\hat{\mathbf{B}} \quad \text{and} \tag{F2}$$

$$\mathbf{\Psi}_2 = \hat{\mathbf{C}}\mathbf{A}\hat{\mathbf{B}} - \mathbf{\Phi}_1\hat{\mathbf{C}}\hat{\mathbf{B}}. \tag{F3}$$

Note from Eq. (F2) that $\mathbf{\Psi}_1$ does not depend on $\mathbf{N}$, (i.e., $\mathbf{\Phi}_1$, $\mathbf{\Phi}_2$). The L_1-norm minimization problem is to find

$$\underset{\mathbf{\Phi}_1,\mathbf{\Phi}_2}{\operatorname{argmin}}\{\|\operatorname{col}([\,\mathbf{\Theta}_\mathbf{y} \quad \mathbf{\Theta}_\mathbf{z}\,])\|_1 + \|\operatorname{col}(\mathbf{\Theta}_{\mathbf{f}1})\|_1\} \tag{F4}$$

subject to the constraints of Eq. (F3) and

$$[\mathbf{\Phi}_2 \quad \mathbf{\Phi}_1]\mathscr{O}_2 = \hat{\mathbf{C}}\mathbf{A}^2 \tag{F5}$$

as well as the auxiliary constraints that link the $\mathbf{\Theta}$ matrices and the coefficient matrices of the VARX model, which follow directly from section 5:

$$\mathbf{\Phi}_1 = \mathbf{\Theta}_\mathbf{y} + \mathbf{\Theta}_\mathbf{z},$$
$$\mathbf{\Phi}_2 = -\mathbf{\Theta}_\mathbf{z}, \quad \text{and}$$
$$\mathbf{\Psi}_2 = \mathbf{\Theta}_{\mathbf{f}1} - \mathbf{\Theta}_\mathbf{z}\mathbf{T}_\mathbf{y}\mathbf{T}_\mathbf{f}^\mathrm{T}. \tag{F6}$$

This L_1-norm minimization can be formulated as a linear programming problem of finding

$$\underset{\mathbf{H},\mathbf{J},\mathbf{P},\mathbf{Q},\mathbf{U},\mathbf{V}}{\arg\min}\left[\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{H}+\mathbf{J}+\mathbf{P}+\mathbf{Q})_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{U}+\mathbf{V})_{ij}\right] \tag{F7}$$

subject to the equality constraints of Eqs. (F5) and (F6) as well as the following inequality constraints:

$$H_{ij} \geq 0;\ J_{ij} \geq 0;\ P_{ij} \geq 0;\ Q_{ij} \geq 0;\ U_{ij} \geq 0;\ V_{ij} \geq 0;$$
$$i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, m, \tag{F8}$$

where subscript ij indicates the ith row and jth column of a matrix.

The optimized $\mathbf{\Theta}$ matrices are then given by

$$\mathbf{\Theta}_\mathbf{y} = \mathbf{H} - \mathbf{J},$$
$$\mathbf{\Theta}_\mathbf{z} = \mathbf{P} - \mathbf{Q}, \quad \text{and}$$
$$\mathbf{\Theta}_{\mathbf{f}1} = \mathbf{U} - \mathbf{V}; \tag{F9}$$

$\mathbf{\Psi}_1$, that is, $\mathbf{\Theta}_\mathbf{f}$, is not affected by the minimization and is given in Eq. (F2).

The equality constraints can be combined in a matrix form for the MATLAB solver as

$$\operatorname{col}\left(\begin{bmatrix}\mathscr{S}_1 & -\mathscr{S}_1 & \mathscr{S}_2 & -\mathscr{S}_2\end{bmatrix}\begin{bmatrix}\mathbf{H}^\mathrm{T}\\ \mathbf{J}^\mathrm{T}\\ \mathbf{P}^\mathrm{T}\\ \mathbf{U}^\mathrm{T}\\ \mathbf{Q}^\mathrm{T}\\ \mathbf{V}^\mathrm{T}\end{bmatrix}\right) = \operatorname{col}\left(\begin{bmatrix}(\hat{\mathbf{C}}\mathbf{A}^2)^\mathrm{T}\\ (\hat{\mathbf{C}}\mathbf{A}\hat{\mathbf{B}})^\mathrm{T}\end{bmatrix}\right), \tag{F10}$$

where we have defined

$$\mathscr{S}_1 \equiv \begin{bmatrix}(\hat{\mathbf{C}}\mathbf{A})^\mathrm{T}\\ (\hat{\mathbf{C}}\hat{\mathbf{B}})^\mathrm{T}\end{bmatrix} \quad \text{and} \quad \mathscr{S}_2 \equiv \begin{bmatrix}-(\hat{\mathbf{C}})^\mathrm{T} & \mathbf{0}\\ -\mathbf{T}_\mathbf{f}\mathbf{T}_\mathbf{y}^\mathrm{T} & \mathbf{I}\end{bmatrix}. \tag{F11}$$

APPENDIX G

Search for the Optimal Orthonormal Transformation

First, within the m-dimensional space, we define the 2D rotational matrix in the plane span by the ith and jth axes:

$$\mathbf{T}(\alpha_r) = \begin{array}{c} \\ \\ i \\ \\ j \\ \\ \end{array}\begin{array}{c}\begin{matrix} & & i & & j & & \end{matrix}\\ \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & & & & & \vdots \\ 0 & & \cos\alpha_r & & \sin\alpha_r & & 0 \\ \vdots & & & \ddots & & & \vdots \\ 0 & & -\sin\alpha_r & & \cos\alpha_r & & 0 \\ \vdots & & & & & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}\end{array} \tag{G1}$$

where $r = (i - 1) \times m + j$ is a plane index and α_r is the rotational angle in the rth plane. This rotational matrix is the identity matrix with the iith, ijth, jith, jjth entries replaced by the four entries involving α_r in Eq. (G1).

The total number of unique pairs of i and j ($i \neq j$), and thus the total number of independent angles, is $m(m - 1)/2$, which are the parameters that we seek to adjust. We can define the transformation matrix for the outputs $\mathbf{y}$ as the product of these 2D rotational matrices:

$$\mathbf{T}_y(\boldsymbol{\alpha}) = \prod_{r=1}^{m(m-1)/2}\mathbf{T}(\alpha_r); \tag{G2}$$

$\boldsymbol{\alpha}$ is the parameter vector that holds all of the $m(m - 1)/2$ angles. Reflections are not included because they do not affect the L_1 norm of the coefficients of the resulting VARX model.

There are no prior reasons that the transformation matrix for the input is the same as that for the output, so we define a separate transformation matrix for the input $\mathbf{f}$ in the same way:

$$\mathbf{T}_f(\boldsymbol{\beta}) = \prod_{r=1}^{m(m-1)/2}\mathbf{T}(\beta_r). \tag{G3}$$

Minimization is done over $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. This is a high-dimensional nonlinear minimization and is solved using the quasi-Newton method, as implemented in MATLAB function *fminunc*. The search is facilitated by the knowledge of the Jacobian, which can be computed using the Lagrange multipliers of the equality constraints from the linear programming step as follows.

From Eq. (G1), $\partial \mathbf{T}(\alpha_r)/\partial\alpha_r$ can be easily computed. We can then compute

$$\frac{\partial \mathbf{T_y}}{\partial \alpha_r} = \left[\prod_{s=1}^{r-1} \mathbf{T}(\alpha_s)\right] \frac{\partial \mathbf{T}(\alpha_r)}{\partial \alpha_r} \left[\prod_{s=r+1}^{m(m-1)/2} \mathbf{T}(\alpha_s)\right] \quad \text{and} \tag{G4}$$

$$\frac{\partial \hat{\mathbf{C}}}{\partial \alpha_r} = \frac{\partial \mathbf{T}_y}{\partial \alpha_r}\mathbf{C}. \tag{G5}$$

To shorten the notation, we define

$$\hat{\mathbf{C}}_{\alpha_r} \equiv \frac{\partial \hat{\mathbf{C}}}{\partial \alpha_r}. \tag{G6}$$

Taking $p = 2$ again as an example, suppose the solution to the L_1-norm minimization problem, given the current $\mathbf{T_y}$ and $\mathbf{T_f}$, is $\mathbf{H}$, $\mathbf{J}$, $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{U}$, and $\mathbf{V}$. The derivatives of the constraints in Eq. (F10) with respect to α_r are then

$$\text{col}\left\{\begin{bmatrix} (\hat{\mathbf{C}}_{\alpha_r}\mathbf{A}^2)^{\mathrm{T}} \\ (\hat{\mathbf{C}}_{\alpha_r}\mathbf{A}\hat{\mathbf{B}})^{\mathrm{T}} \end{bmatrix} - \begin{bmatrix} \mathscr{S}_{1,\alpha_r} & -\mathscr{S}_{1,\alpha_r} & \mathscr{S}_{2,\alpha_r} & -\mathscr{S}_{2,\alpha_r} \end{bmatrix} \begin{bmatrix} \mathbf{H}^{\mathrm{T}} \\ \mathbf{J}^{\mathrm{T}} \\ \mathbf{P}^{\mathrm{T}} \\ \mathbf{Q}^{\mathrm{T}} \end{bmatrix}\right\}, \tag{G7}$$

with

$$\mathscr{S}_{1,\alpha_r} \equiv \begin{bmatrix} (\hat{\mathbf{C}}_{\alpha_r}\mathbf{A})^{\mathrm{T}} \\ (\hat{\mathbf{C}}_{\alpha_r}\hat{\mathbf{B}})^{\mathrm{T}} \end{bmatrix} \quad \text{and} \quad \mathscr{S}_{2,\alpha_r} \equiv \begin{bmatrix} -(\hat{\mathbf{C}}_{\alpha_r})^{\mathrm{T}} \\ -\left(\dfrac{\partial \mathbf{T_y}}{\partial \alpha_r}\mathbf{T_f}^{\mathrm{T}}\right)^{\mathrm{T}} \end{bmatrix}. \tag{G8}$$

The dot product of this change in the constraints with the Lagrange multipliers of the corresponding equality constraints from the linear programming step gives how much the minimum L_1 norm of col($[\boldsymbol{\Theta}_\mathbf{y}\boldsymbol{\Theta}_\mathbf{z}\boldsymbol{\Theta}_{\mathbf{f}1}]$) decreases as α_r increases. The $\boldsymbol{\Theta}_\mathbf{f}$ can be included by adding the term

$$-\sum_{i,j} (\hat{\mathbf{C}}_{\alpha_r})_{ij} \,\text{sgn}[(\hat{\mathbf{C}}\hat{\mathbf{B}})_{ij}], \tag{G9}$$

where sgn$(\cdot)$ is the sign function.

For β_r, the expressions equivalent to Eqs. (G4) and (G5) are

$$\frac{\partial \mathbf{T_f}}{\partial \beta_r} = \left[\prod_{s=1}^{r-1} \mathbf{T}(\beta_s)\right] \frac{\partial \mathbf{T}(\beta_r)}{\partial \beta_r} \left[\prod_{s=r+1}^{m(m-1)/2} \mathbf{T}(\beta_s)\right] \quad \text{and}$$

$$\frac{\partial \hat{\mathbf{B}}}{\partial \beta_r} = \mathbf{B}\left(\frac{\partial \mathbf{T_f}}{\partial \beta_r}\right)^{\mathrm{T}}, \tag{G10}$$

and the derivatives of the constraints with respect to β_r are

$$\text{col}\left\{\left(\hat{\mathbf{C}}\mathbf{A}\frac{\partial \hat{\mathbf{B}}}{\partial \beta_r}\right)^{\mathrm{T}} - \begin{bmatrix} \mathscr{S}_{1,\beta_r} & -\mathscr{S}_{1,\beta_r} & \mathscr{S}_{2,\beta_r} & -\mathscr{S}_{2,\beta_r} \end{bmatrix} \begin{bmatrix} \mathbf{H}^{\mathrm{T}} \\ \mathbf{J}^{\mathrm{T}} \\ \mathbf{P}^{\mathrm{T}} \\ \mathbf{Q}^{\mathrm{T}} \end{bmatrix}\right\}, \tag{G11}$$

with

$$\mathscr{S}_{1,\beta_r} \equiv \left(\hat{\mathbf{C}}\frac{\partial \hat{\mathbf{B}}}{\partial \beta_r}\right)^{\mathrm{T}} \quad \text{and} \quad \mathscr{S}_{2,\beta_r} \equiv -\left(\frac{\partial \mathbf{T_f}}{\partial \beta_r}\right)\mathbf{T_y}^{\mathrm{T}}. \tag{G12}$$

Similarly, the dot product of Eq. (G11) with the Lagrange multipliers of the corresponding equality constraints from the linear programming step gives how much the minimum L_1 norm of col($[\boldsymbol{\Theta}_\mathbf{y}\boldsymbol{\Theta}_\mathbf{z}\boldsymbol{\Theta}_{\mathbf{f}1}]$) decreases as β_r increases. The $\boldsymbol{\Theta}_\mathbf{f}$ can be included by adding the term

$$-\sum_{i,j}\left(\frac{\partial \hat{\mathbf{B}}}{\partial \beta_r}\right)_{ij} \text{sgn}[(\hat{\mathbf{C}}\hat{\mathbf{B}})_{ij}]. \tag{G13}$$

Computing these variations over all α_r and β_r gives the negative of the Jacobian for the quasi-Newton method used in MATLAB function fminunc. We initialize the nonlinear minimization with all values of α_r and β_r set to zero. While we cannot prove the minima found are the global minima, extensive experimentation with randomized initial values of α_r and β_r did not produce lower minima.

APPENDIX H

Additional Considerations

With CVA weighting and prediction focus, for $n \leq 40$, all optimized $\boldsymbol{\Theta}$ matrices except $\boldsymbol{\Theta}_\mathbf{y}$ and $\boldsymbol{\Theta}_\mathbf{f}$ are zero. For $n = 40$, $\boldsymbol{\Theta}_\mathbf{y}$ and $\boldsymbol{\Theta}_\mathbf{f}$, both 40×40 matrices, are full rank. This indicates that the identified model is the optimal model (measured in L_2 norm) without accounting for convection memory. This facilitates the interpretation of convection memory, because further increases in the model order n incrementally add contributions purely from past inputs and outputs. Furthermore, for $n > 40$, $\boldsymbol{\Theta}_\mathbf{y}$ and $\boldsymbol{\Theta}_\mathbf{f}$ are always full rank. Therefore, for $n > 40$, we exclude $\boldsymbol{\Theta}_\mathbf{y}$ and $\boldsymbol{\Theta}_\mathbf{f}$ in the L_1-norm minimization procedure to focus on enhancing the sparsity of the other $\boldsymbol{\Theta}$ matrices. This is found to produce cleaner results, which we present in section 5b. The exclusion of $\boldsymbol{\Theta}_\mathbf{y}$ and $\boldsymbol{\Theta}_\mathbf{f}$ is done by setting weights of $\boldsymbol{\Theta}_\mathbf{y}$ in Eq. (F4) to zero and removing contributions from Eqs. (G9) and (G13) when computing the Jacobian.

With MOESP weighting and simulation focus, $\boldsymbol{\Theta}$ matrices other than $\boldsymbol{\Theta}_\mathbf{y}$ and $\boldsymbol{\Theta}_\mathbf{f}$ can be nonzero even when $n \leq 40$. This is because MOESP weighting with simulation focus does not weigh errors by the inverse of the error covariance and eliminates modes in the current inputs and outputs that contribute little to the L_2 norm of the residual in favor of past inputs and outputs that contribute more. The eliminated modes in the current inputs and outputs tend to have smaller

error covariance. When CVA weighting with prediction focus normalizes errors by the error covariance, the importance of those modes is enhanced such that they are retained instead of past inputs and outputs. As seen in Fig. 1b, when the model order n is less than 43, MOESP gives substantially better performance than CVA for a given model order. However, in the MOESP case, increases in the model order add contributions from both past and present inputs and outputs, complicating the interpretation. Therefore, we restrict ourselves to CVA weighting with prediction focus in section 5 when discussing interpretations of convection memory.

REFERENCES

Arakawa, A., and W. H. Schubert, 1974: Interaction of a cumulus cloud ensemble with the large-scale environment, Part I. *J. Atmos. Sci.*, **31**, 674–701, https://doi.org/10.1175/1520-0469(1974)031<0674:IOACCE>2.0.CO;2.

Arnold, N. P., and D. A. Randall, 2015: Global-scale convective aggregation: Implications for the Madden-Julian oscillation. *J. Adv. Model. Earth Syst.*, **7**, 1499–1518, https://doi.org/10.1002/2015MS000498.

Athanasopoulos, G., D. S. Poskitt, and F. Vahid, 2012: Two canonical VARMA forms: Scalar component models vis-à-vis the echelon form. *Econom. Rev.*, **31**, 60–83, https://doi.org/10.1080/07474938.2011.607088.

Brenowitz, N. D., and C. S. Bretherton, 2018: Prognostic validation of a neural network unified physics parameterization. *Geophys. Res. Lett.*, **45**, 6289–6298, https://doi.org/10.1029/2018GL078510.

Candès, E. J., J. K. Romberg, and T. Tao, 2006: Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.*, **59**, 1207–1223, https://doi.org/10.1002/cpa.20124.

Donoho, D. L., 2006: For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Commun. Pure Appl. Math.*, **59**, 797–829, https://doi.org/10.1002/cpa.20132.

Emanuel, K. A., J. D. Neelin, and C. S. Bretherton, 1994: On large-scale circulations in convecting atmospheres. *Quart. J. Roy. Meteor. Soc.*, **120**, 1111–1143, https://doi.org/10.1002/qj.49712051902.

Golaz, J.-C., V. E. Larson, and W. R. Cotton, 2002a: A PDF-based model for boundary layer clouds. Part I: Method and model description. *J. Atmos. Sci.*, **59**, 3540–3551, https://doi.org/10.1175/1520-0469(2002)059<3540:APBMFB>2.0.CO;2.

——, ——, and ——, 2002b: A PDF-based model for boundary layer clouds. Part II: Model results. *J. Atmos. Sci.*, **59**, 3552–3571, https://doi.org/10.1175/1520-0469(2002)059<3552:APBMFB>2.0.CO;2.

Han, Y., G. J. Zhang, X. Huang, and Y. Wang, 2020: A moist physics parameterization based on deep learning. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002076, https://doi.org/10.1029/2020MS002076.

——, ——, and Y. Wang, 2023: An ensemble of neural networks for moist physics processes, its generalizability and stable integration. *J. Adv. Model. Earth Syst.*, **15**, e2022MS003508, https://doi.org/10.1029/2022MS003508.

Hassanzadeh, P., and Z. Kuang, 2016: The linear response function of an idealized atmosphere. Part I: Construction using Green's functions and applications. *J. Atmos. Sci.*, **73**, 3423–3439, https://doi.org/10.1175/JAS-D-15-0338.1.

Jones, T. R., and D. A. Randall, 2011: Quantifying the limits of convective parameterizations. *J. Geophys. Res.*, **116**, D08210, https://doi.org/10.1029/2010JD014913.

Kalman, R. E., 1960: A new approach to linear filtering and prediction problems. *J. Basic Eng.*, **82**, 35–45, https://doi.org/10.1115/1.3662552.

Khairoutdinov, M. F., and D. A. Randall, 2003: Cloud resolving modeling of the ARM summer 1997 IOP: Model formulation, results, uncertainties, and sensitivities. *J. Atmos. Sci.*, **60**, 607–625, https://doi.org/10.1175/1520-0469(2003)060<0607:CRMOTA>2.0.CO;2.

——, C. DeMott, and D. Randall, 2008: Evaluation of the simulated interannual and subseasonal variability in an AMIP-style simulation using the CSU Multiscale Modeling Framework. *J. Climate*, **21**, 413–431, https://doi.org/10.1175/2007JCLI1630.1.

Kuang, Z., 2008a: Modeling the interaction between cumulus convection and linear gravity waves using a limited-domain cloud system–resolving model. *J. Atmos. Sci.*, **65**, 576–591, https://doi.org/10.1175/2007JAS2399.1.

——, 2008b: A moisture-stratiform instability for convectively coupled waves. *J. Atmos. Sci.*, **65**, 834–854, https://doi.org/10.1175/2007JAS2444.1.

——, 2010: Linear response functions of a cumulus ensemble to temperature and moisture perturbations and implications for the dynamics of convectively coupled waves. *J. Atmos. Sci.*, **67**, 941–962, https://doi.org/10.1175/2009JAS3260.1.

——, 2012: Weakly forced mock Walker cells. *J. Atmos. Sci.*, **69**, 2759–2786, https://doi.org/10.1175/JAS-D-11-0307.1.

——, 2018: Linear stability of moist convecting atmospheres. Part I: From linear response functions to a simple model and applications to convectively coupled waves. *J. Atmos. Sci.*, **75**, 2889–2907, https://doi.org/10.1175/JAS-D-18-0092.1.

Larimore, W. E., 1990: Canonical variate analysis in identification, filtering, and adaptive-control. *29th IEEE Conf. on Decision and Control*, Honolulu, HI, IEEE, 596–604, https://doi.org/10.1109/CDC.1990.203665.

Ljung, L., 1999: *System Identification: Theory for the User*. 2nd ed. Prentice Hall, 631 pp.

Luenberger, D. G., 1967: Canonical forms for linear multivariable systems. *IEEE Trans. Autom. Control*, **12**, 290–293, https://doi.org/10.1109/TAC.1967.1098584.

Mapes, B. E., 2000: Convective inhibition, subgrid-scale triggering energy, and stratiform instability in a toy tropical wave model. *J. Atmos. Sci.*, **57**, 1515–1535, https://doi.org/10.1175/1520-0469(2000)057<1515:CISSTE>2.0.CO;2.

——, and R. Neale, 2011: Parameterizing convective organization to escape the entrainment dilemma. *J. Adv. Model. Earth Syst.*, **3**, M06004, https://doi.org/10.1029/2011MS000042.

Neelin, J. D., O. Peters, J. W.-B. Lin, K. Hales, and C. E. Holloway, 2008: Rethinking convective quasi-equilibrium: Observational constraints for stochastic convective schemes in climate models. *Philos. Trans. Roy. Soc.*, **A366**, 2581–2604, https://doi.org/10.1098/rsta.2008.0056.

O'Gorman, P. A., and J. G. Dwyer, 2018: Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *J. Adv. Model. Earth Syst.*, **10**, 2548–2563, https://doi.org/10.1029/2018MS001351.

Palmer, T. N., 2019: Stochastic weather and climate models. *Nat. Rev. Phys.*, **1**, 463–471, https://doi.org/10.1038/s42254-019-0062-2.

Pauluis, O., and S. Garner, 2006: Sensitivity of radiative–convective equilibrium simulations to horizontal resolution. *J. Atmos. Sci.*, **63**, 1910–1923, https://doi.org/10.1175/JAS3705.1.

Phan, M. Q., R. K. Lim, and R. W. Longman, 1998: Unifying input-output and state-space perspectives of predictive control. Princeton University Dept. of Mechanical and Aerospace Engineering Tech. Rep. 3044, 35 pp.

Pintelon, R., and J. Schoukens, 2012: *System Identification: A Frequency Domain Approach*. 2nd ed. Wiley, 788 pp.

Rasp, S., M. S. Pritchard, and P. Gentine, 2018: Deep learning to represent subgrid processes in climate models. *Proc. Natl. Acad. Sci. USA*, **115**, 9684–9689, https://doi.org/10.1073/pnas.1810286115.

Schoukens, J., M. Vaes, and R. Pintelon, 2016: Linear system identification in a nonlinear setting: Nonparametric analysis of the nonlinear distortions and their impact on the best linear approximation. *IEEE Control Syst. Mag.*, **36**, 38–69, https://doi.org/10.1109/MCS.2016.2535918.

Shamekh, S., K. D. Lamb, Y. Huang, and P. Gentine, 2023: Implicit learning of convective organization explains precipitation stochasticity. *Proc. Natl. Acad. Sci. USA*, **120**, e2216158120, https://doi.org/10.1073/pnas.2216158120.

Tiao, G. C., and R. S. Tsay, 1989: Model specification in multivariate time series. *J. Roy. Stat. Soc.*, **51B**, 157–195, https://doi.org/10.1111/j.2517-6161.1989.tb01756.x.

Verhaegen, M., 1994: Identification of the deterministic part of MIMO state-space models given in innovations form from input-output data. *Automatica*, **30**, 61–74, https://doi.org/10.1016/0005-1098(94)90229-1.

Wheeler, M., and G. N. Kiladis, 1999: Convectively coupled equatorial waves: Analysis of clouds and temperature in the wavenumber–frequency domain. *J. Atmos. Sci.*, **56**, 374–399, https://doi.org/10.1175/1520-0469(1999)056<0374:CCEWAO>2.0.CO;2.

Yuval, J., P. A. O'Gorman, and C. N. Hill, 2021: Use of neural networks for stable, accurate and physically consistent parameterization of subgrid atmospheric processes with good performance at reduced precision. *Geophys. Res. Lett.*, **48**, e2020GL091363, https://doi.org/10.1029/2020GL091363.