Energetic Consistency and Coupling of the Mean and Covariance Dynamics[1]

Stephen E. Cohn
Global Modeling and Assimilation Office
NASA Goddard Space Flight Center
Greenbelt, Maryland 20771

Draft of March 29, 2008

[1]prepared for *Handbook of Numerical Analysis: Special Volume on Computational Methods for the Ocean and the Atmosphere*, R. Temam and J. Tribbia, eds., Elsevier

1 Introduction

The dynamical state of the ocean and atmosphere is taken to be a large-dimensional random vector in a range of large-scale computational applications, including data assimilation, ensemble prediction, sensitivity analysis, and predictability studies. In each of these applications, numerical evolution of the covariance matrix of the random state plays a central role, because this matrix is used to quantify uncertainty in the state of the dynamical system. Since atmospheric and ocean dynamics are nonlinear, there is no closed evolution equation for the covariance matrix, nor for the mean state. Therefore approximate evolution equations must be used.

This article studies theoretical properties of the evolution equations for the mean state and covariance matrix that arise in the second-moment closure approximation (third- and higher-order moment discard). This approximation was introduced by EPSTEIN [1969] in an early effort to introduce a stochastic element into deterministic weather forecasting, and was studied further by FLEMING [1971a,b], EPSTEIN and PITCHER [1972], and PITCHER [1977], also in the context of atmospheric predictability. It has since fallen into disuse, with a simpler one being used in current large-scale applications. The theoretical results of this article make a case that this approximation should be reconsidered for use in large-scale applications, however, because the second-moment closure equations possess a property of energetic consistency that the approximate equations now in common use do not possess. A number of properties of solutions of the second-moment closure equations that result from this energetic consistency will be established.

Suppose the dynamics of the state $\mathbf{s} \in \mathbb{R}^N$ are given by a system of nonlinear ordinary differential equations,

$$\frac{d\mathbf{s}}{dt} + \mathbf{f}(\mathbf{s}, t) = \mathbf{0}, \tag{1}$$

where t is time, $\mathbf{f} : \mathcal{S} \times \mathcal{T} \to \mathbb{R}^N$, $\mathcal{S} \subseteq \mathbb{R}^N$ is a state space appropriate for Eq. (1), and $\mathcal{T} = [t_0, T]$ is a closed time interval. The initial condition $\mathbf{s}_{t_0}$ is taken to be a random state, $\mathbf{s}_{t_0} \in \mathcal{S}$ with probability one, so that the problem to be solved is the stochastic initial-value problem for Eq. (1). Technical assumptions on $\mathbf{f}$ are stated in Secs. 2 and 5. In addition, it will be assumed that the dynamics are conservative, and a nonlinear transformation is introduced in Sec. 3 to ensure that the total energy conserved by solutions of Eq. (1), stochastic or deterministic, is just $E = \frac{1}{2}\|\mathbf{s}\|^2$, where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^N$.

A simple sufficient condition, which is natural for conservative dynamics, under which the stochastic initial-value problem for Eq. (1) is well-posed, is stated in Sec. 4. Under this condition, the solution of the stochastic initial-value problem defines a second-order stochastic process, that is, one that has a mean and covariance matrix at each time t in a closed time interval. Furthermore, it follows immediately from conservation of total energy E for this process that

$$\frac{d(\|\overline{\mathbf{s}}\|^2 + \operatorname{tr} \mathbf{P})}{dt} = 0, \tag{2}$$

where $\overline{\mathbf{s}} = \overline{\mathbf{s}}_t \in \mathcal{S}$ is the mean state of this process, $\mathbf{P} = \mathbf{P}_t \in \mathbb{R}^{N\times N}$ is the covariance matrix of the process, and $\operatorname{tr}\mathbf{P}$ is the trace, or sum of the diagonal elements, of $\mathbf{P}$. This means that the uncertainty in the random state $\mathbf{s}$, as measured by the total variance $V = \operatorname{tr}\mathbf{P}$, can increase (decrease) only as a result of extracting energy from (inserting energy into) the mean state $\overline{\mathbf{s}}$, with the change in total variance balanced exactly by twice the change in total energy $\frac{1}{2}\|\overline{\mathbf{s}}\|^2$ of the mean state. The mean state and covariance matrix cannot be calculated from Eq. (1) without approximation, however, unless $\mathbf{f}$ is linear in $\mathbf{s}$, and one would like to develop approximate evolution equations for $\overline{\mathbf{s}}$ and $\mathbf{P}$ that in the nonlinear case at least preserve this basic conservation property. A closed system of ordinary differential equations for the mean and covariance matrix whose solutions satisfy Eq. (2) will be said to be *energetically consistent*, after FLEMING [1971a, p. 872].

The second-moment closure equations for approximate evolution of $\overline{\mathbf{s}}$ and $\mathbf{P}$ are the closed, nonlinearly coupled differential equations

$$\frac{d\overline{\mathbf{s}}}{dt} + \mathbf{f}(\overline{\mathbf{s}}, t) + \tfrac{1}{2}\sum_j \sum_k \frac{\partial^2 \mathbf{f}(\overline{\mathbf{s}}, t)}{\partial s_j \partial s_k} P_{jk} = \mathbf{0}, \tag{3}$$

$$\frac{d\mathbf{P}}{dt} + \mathbf{F}(\overline{\mathbf{s}}, t)\mathbf{P} + \mathbf{P}\mathbf{F}^T(\overline{\mathbf{s}}, t) = \mathbf{0}, \tag{4}$$

where P_{jk} is the $(j,k)^{th}$ element of the covariance matrix $\mathbf{P}$ whose evolution is described by Eq. (4), $\mathbf{F} = \partial\mathbf{f}/\partial\mathbf{s}$ is the Jacobian matrix of $\mathbf{f}$, and the superscript T denotes transposition. The evolution of $\mathbf{P}$ in Eq. (4) depends on that of the mean state $\overline{\mathbf{s}}$ given by Eq. (3), through the dependence of the Jacobian matrix on the mean state, and the evolution of the mean state also depends on that of the covariance matrix, through the double-summation term in Eq. (3). If $\mathbf{f}(\mathbf{s}, t)$ is linear in $\mathbf{s}$, then the second partial derivatives of $\mathbf{f}$ with respect to the state variables all vanish, so that the double-summation term vanishes. Hence this term is called the nonlinear coupling term. Equations (3) and (4) are to be solved together for initial conditions $\overline{\mathbf{s}}_{t_0}$ and $\mathbf{P}_{t_0}$, with $\overline{\mathbf{s}}_{t_0}$ being the mean of the random initial condition $\mathbf{s}_{t_0}$ for Eq. (1) and $\mathbf{P}_{t_0}$ being the covariance matrix of $\mathbf{s}_{t_0}$.

Conditions under which this initial-value problem is well-posed on a closed time interval are given in Sec. 5, where a stochastic process having the solution $(\overline{\mathbf{s}}, \mathbf{P})$ of the initial-value problem as its mean and covariance matrix is also defined. It is shown in Sec. 6 that the solution satisfies Eq. (2). Thus the second-moment closure equations are energetically consistent. For quadratically nonlinear $\mathbf{f}$, energetic consistency of the second-moment closure equations was established by EPSTEIN [1969] and studied in detail by FLEMING [1971a], who also established energetic consistency of the third-moment closure equations for quadratically nonlinear $\mathbf{f}$. The energetic consistency result established in the present article holds for general $\mathbf{f}$.

The derivation of Eq. (2) given in Sec. 6 for the second-moment closure equations shows that the exchange of energy between the mean state and the

stochastic perturbations, which balances exactly, occurs solely through the nonlinear coupling term in Eq. (3) and the symmetric part $\mathbf{F}^s = \frac{1}{2}(\mathbf{F} + \mathbf{F}^T)$ of the Jacobian matrix in Eq. (4). In other words, rewriting Eq. (4) as

$$\frac{d\mathbf{P}}{dt} + \mathbf{F}^a(\overline{\mathbf{s}}, t)\mathbf{P} - \mathbf{P}\mathbf{F}^a(\overline{\mathbf{s}}, t) + \mathbf{F}^s(\overline{\mathbf{s}}, t)\mathbf{P} + \mathbf{P}\mathbf{F}^s(\overline{\mathbf{s}}, t) = \mathbf{0}, \tag{5}$$

where $\mathbf{F}^a = \frac{1}{2}(\mathbf{F} - \mathbf{F}^T)$ is the anti-symmetric (skew-symmetric) part of $\mathbf{F}$, one has immediately that

$$\frac{d \operatorname{tr} \mathbf{P}}{dt} + 2 \operatorname{tr} \mathbf{F}^s(\overline{\mathbf{s}}, t)\mathbf{P} = 0,$$

and it is shown in Sec. 6 that Eq. (3) gives

$$\frac{d\|\overline{\mathbf{s}}\|^2}{dt} - 2 \operatorname{tr} \mathbf{F}^s(\overline{\mathbf{s}}, t)\mathbf{P} = 0,$$

with contribution only from the nonlinear coupling term, not from the term $\mathbf{f}(\overline{\mathbf{s}}, t)$ in Eq. (3). Equation (2) then follows.

In case $\mathbf{f}(\mathbf{s}, t)$ is linear in $\mathbf{s}$, then not only does the nonlinear coupling term vanish, but for conservative dynamics $\mathbf{F}^s$ vanishes as well, so that Eqs. (3) and (4) become simply

$$\frac{d\overline{\mathbf{s}}}{dt} + \mathbf{f}(\overline{\mathbf{s}}, t) = \mathbf{0}, \tag{6}$$

$$\frac{d\mathbf{P}}{dt} + \mathbf{F}^a\mathbf{P} - \mathbf{P}\mathbf{F}^a = \mathbf{0}, \tag{7}$$

with $\mathbf{F}^a$ independent of $\overline{\mathbf{s}}$ by linearity. Thus the effect of nonlinearity, to second-moment closure, is to introduce the nonlinear coupling term and the terms $\mathbf{F}^s(\overline{\mathbf{s}}, t)\mathbf{P} + \mathbf{P}\mathbf{F}^s(\overline{\mathbf{s}}, t)$, and these terms together are in energetic balance. Nonlinearity also introduces dependence of $\mathbf{F}^a$ on the mean state, but Eq. (7) is energetically neutral, satisfying $d \operatorname{tr} \mathbf{P}/dt = 0$, regardless of any dependence of $\mathbf{F}^a$ on the mean state.

The approximation now widely used in large-scale atmospheric and oceanic applications is to retain Eq. (4) as it stands, but to neglect the nonlinear coupling term. This is the approximation made for instance in four-dimensional variational data assimilation (TALAGRAND and COURTIER [1987]; COURTIER and TALAGRAND [1987]; THÉPAUT et al. [1996]) and in a variety of algorithms based on singular vector calculations (e.g. BUIZZA and PALMER [1995]; MOLTENI et al. [1996]; MOORE and KLEEMAN [1997]). This approximation is convenient for computations, because the mean state can then be evolved independently of the covariance matrix. Neglecting the nonlinear coupling term, however, destroys energetic consistency. Furthermore, the nonlinear coupling term and the terms in the covariance evolution equation all have formally the same order of magnitude, since all are linear in the covariance matrix. Moreover, in the sense of contribution to the total variance, the terms retained in the covariance evolution equation that arise from nonlinearity, $\mathbf{F}^s\mathbf{P}+\mathbf{P}\mathbf{F}^s$, have precisely the same magnitude, and opposite sign, as that of the nonlinear coupling term which is neglected.

The role of $\mathbf{F}^s$ in the energetic coupling of the second-moment closure equations is studied further in Secs. 7 and 8. It is shown in Sec. 8 that if $\mathbf{f}$ is *genuinely nonlinear*, as defined there, then $\mathbf{F}^s$ has at least one positive and one negative eigenvalue. This means that when the dynamics are genuinely nonlinear, there is always a direction in state space in which uncertainty decays, as well as a direction in which uncertainty grows. One implication is that if $\mathbf{f}$ is genuinely nonlinear, then neglect of the nonlinear coupling term can lead either to increase or decrease of the perceived uncertainty.

When the nonlinear coupling term is neglected, Eq. (4) is a linear equation to be solved once the mean state has been calculated, and so its solution can be expressed in the form

$$\mathbf{P}_t = \mathbf{M}_{t,t_0} \mathbf{P}_{t_0} \mathbf{M}_{t,t_0}^T, \tag{8}$$

where $\mathbf{M}_{t,t_0}$ is the fundamental matrix (alternatively, solution operator, or tangent linear propagator) of the perturbation dynamics corresponding to Eq. (1) linearized about the mean state. This expression is particularly convenient for singular-value calculations in large-scale applications.

When the nonlinear coupling term is retained, the solution of Eq. (4) can still be expressed in the form (8), but for a matrix $\mathbf{M}_{t,t_0}$ that itself depends on the covariance matrix. In Sec. 9, Eq. (2) is used to establish simple time-independent upper bounds for $\operatorname{tr} \mathbf{P}_t / \operatorname{tr} \mathbf{P}_{t_0}$, which hold also for the covariance matrix of the original stochastic process. When the nonlinear coupling term is neglected, the largest singular value of $\mathbf{M}_{t,t_0}$ is the least upper bound for $\operatorname{tr} \mathbf{P}_t / \operatorname{tr} \mathbf{P}_{t_0}$, but in general there is no time-independent upper bound since Eq. (2) does not hold.

A minimum requirement for solutions of Eqs. (3) and (4) to approximate the mean and covariance matrix of solutions of the stochastic initial-value problem for Eq. (1) is for $\overline{\mathbf{s}}_{t_0}$ and $\mathbf{P}_{t_0}$ to be the mean and covariance matrix of some random state, $\mathbf{s}_{t_0} \in \mathcal{S}$ with probability one. Because ocean and atmospheric dynamics contain state variables that are constrained to be positive, such as mass and temperature variables, or to satisfy other constraints, this minimum requirement implies that $\mathbf{P}_{t_0}$ cannot be chosen independently of $\overline{\mathbf{s}}_{t_0}$ in general. Thus the initial-value problem for Eqs. (3) and (4) must be posed with some care. Toward this end, the deterministic initial-value problem for Eq. (1) is reviewed briefly in Sec. 2, and the stochastic initial-value problem for Eq. (1) is described in some detail in Sec. 4. The initial-value problem for Eqs. (3) and (4) is posed in Sec. 5, and restrictions on the initial covariance matrix are illustrated there using a spatially discretized version of the one-dimensional shallow-water equations.

Brief concluding remarks are given in Sec. 10.

2 The Deterministic Initial-Value Problem

Suppose that the evolution of a real vector $\mathbf{s} \in \mathbb{R}^N$ is governed by a nonlinear system of ordinary differential equations,

$$\frac{d\mathbf{s}}{dt} + \mathbf{f}(\mathbf{s}) = \mathbf{0}, \tag{9}$$

where $\mathbf{f}(\mathbf{s}) = \mathbf{f}(\mathbf{s}, t)$ may depend explicitly on time t. Suppose also that $\mathbf{f} : \mathcal{S} \times \mathcal{T} \to \mathbb{R}^N$, where $\mathcal{S} \subseteq \mathbb{R}^N$ is a convex open set, that is, an open set in $\mathbb{R}^N$ such that the line segment between any two points in $\mathcal{S}$ lies entirely in $\mathcal{S}$, and where $\mathcal{T} = [t_0, T]$ is a fixed, closed time interval; $T - t_0$ is finite but may be arbitrarily large. The open set $\mathcal{S}$ is called the state space, an element $\mathbf{s} \in \mathcal{S}$ is called a state, or state vector, and the N components of a state are the state variables. If the actual system under consideration is complex, Eq. (9) is obtained by separation into real and imaginary parts.

This section provides a brief review of the deterministic initial-value problem for Eq. (9), in which one is supposed to find, for each initial state $\mathbf{s}_{t_0} \in \mathcal{S}$, a solution $\mathbf{s} = \mathbf{s}(t) \in \mathcal{S}$ that satisfies $\mathbf{s}(t_0) = \mathbf{s}_{t_0}$. Sufficiently strong hypotheses will be imposed on $\mathbf{f}$ to guarantee existence and uniqueness of solutions on a (possibly short) half-open time interval $\mathcal{T}_* = [t_0, T_*) \subset \mathcal{T}$, with T_* depending in general on $\mathbf{s}_{t_0}$, and also to guarantee that the solution is in the space $C^1(\mathcal{T}_*)$ of functions with one continuous time derivative on $\mathcal{T}_*$. The corresponding stochastic initial-value problem, in which $\mathbf{s}_{t_0}$ will depend on a probability variable, is considered in Sec. 4. It should be noted here that Assumption 2 below is stronger than necessary for establishing uniqueness of solutions of Eq. (9), but that this assumption will be required later, in Sec. 5, to ensure that solutions of the covariance evolution equation have one continuous time derivative.

Assumption 1 $\mathbf{f} \in C(\mathcal{S} \times \mathcal{T})$, *the space of continuous functions on* $\mathcal{S} \times \mathcal{T}$.

Assumption 2 $\mathbf{F} = \mathbf{F}(\mathbf{s}, t) \in C(\mathcal{S} \times \mathcal{T})$, *where* $\mathbf{F} = \partial\mathbf{f}/\partial\mathbf{s}$ *denotes the* $N \times N$ *Jacobian matrix of* $\mathbf{f}$, *whose* $(j,k)^{th}$ *element is given by*

$$F_{jk}(\mathbf{s}, t) = \frac{\partial f_j(\mathbf{s}, t)}{\partial s_k}.$$

That is, $F_{jk} \in C(\mathcal{S} \times \mathcal{T})$ *for* $j, k = 1, \ldots, N$.

Assumption 1 guarantees that for each $\mathbf{s}_{t_0} \in \mathcal{S}$, there exists a time interval $\mathcal{T}_* = [t_0, T_*)$, with $T_* = T_*(\mathbf{s}_{t_0}) \leq T$, and at least one solution $\mathbf{s}(t) \in \mathcal{S}$ for all $t \in \mathcal{T}_*$, such that $\mathbf{s}(t_0) = \mathbf{s}_{t_0}$. It also guarantees that every such solution is in $C^1(\mathcal{T}_*)$. Furthermore, existence of a solution ceases only if it hits the boundary of $\mathcal{S}$ (e.g. CODDINGTON and LEVINSON [1955, Ch. 1, Thm. 4.1]), where $\mathbf{f}$ may not even be defined. Assumption 2, along with the fact that $\mathcal{S}$ was taken to be convex, implies that $\mathbf{f}$ satisfies on $\mathcal{S} \times \mathcal{T}$ a Lipschitz condition in $\mathbf{s}$, uniformly in t, and this in turn guarantees that there is at most one solution on any time interval.

In particular, if the state space $\mathcal{S}$ is all of $\mathbb{R}^N$, then $\mathcal{S}$ has no boundary, and so for each $\mathbf{s}_{t_0} \in \mathcal{S}$ there exists a unique solution $\mathbf{s}(t) \in \mathcal{S}$ over the full time interval $\mathcal{T}$, with $\mathbf{s}(t_0) = \mathbf{s}_{t_0}$, and this solution is in $C^1(\mathcal{T})$. More generally, the same is true if the state space is not all of $\mathbb{R}^N$, that is, if Assumptions 1 and 2 hold only for a proper subset $\mathcal{S}$ of $\mathbb{R}^N$, provided the dynamics (9) are such that, for some choice of the state space, no solution can ever hit the boundary of $\mathcal{S}$ by time T. It is often the case that the physical problem at hand dictates that the state space cannot be all of $\mathbb{R}^N$, but that there is a choice of state space for which all solutions exist in $\mathcal{S}$ over the full time interval $\mathcal{T}$, as discussed further in Sec. 3.

Hereafter, the unique solution $\mathbf{s}(t) \in \mathcal{S}$ of Eq. (9) on some time interval $\mathcal{T}_* = \mathcal{T}_*(\mathbf{s}_{t_0}) \subset \mathcal{T}$, such that $\mathbf{s}(t_0) = \mathbf{s}_{t_0}$, or on all of $\mathcal{T}$ in case it exists for all time $t \in \mathcal{T}$, will be denoted by $\mathbf{s}_t$. The continuous path through state space traced by $\mathbf{s}_t$ as time progresses is called the trajectory corresponding to $\mathbf{s}_{t_0}$. The bold letter $\mathbf{s}$ without a subscript will usually denote an arbitrary point in the state space, or in $\mathbb{R}^N$.

An implication of Assumption 2 beyond uniqueness of solutions, which is important for application to the covariance evolution equation, is that solutions depend continuously on parameters such as initial conditions. Regarding each trajectory $\mathbf{s}_t$ as a function of its initial point $\mathbf{s}_{t_0}$, one finds that the $N \times N$ matrix $\mathbf{M} = \mathbf{M}(t) = \partial \mathbf{s}_t / \partial \mathbf{s}_{t_0}$, whose $(j,k)^{th}$ element is given by

$$M_{jk} = \frac{\partial (\mathbf{s}_t)_j}{\partial (\mathbf{s}_{t_0})_k},$$

satisfies the simple *linear* equation

$$\frac{d\mathbf{M}}{dt} + \mathbf{F}(\mathbf{s}_t, t)\mathbf{M} = \mathbf{0}, \tag{10}$$

with initial condition $\mathbf{M} = \mathbf{I}$, the $N \times N$ identity matrix. Although this equation is linear, it is coupled with Eq. (9) through the dependence of the Jacobian matrix $\mathbf{F}$ on the trajectory $\mathbf{s}_t$. Equation (10) is obtained by differentiating Eq. (9) with respect to $\mathbf{s}_{t_0}$ and applying the chain rule. That there exists a unique solution $\mathbf{M} \in \mathbb{R}^{N \times N}$ of Eq. (10), in fact with one continuous time derivative, for as long as the trajectory $\mathbf{s}_t$ exists in $\mathcal{S}$, is guaranteed by Assumption 2 and the linearity of Eq. (10). The dependence of the solution $\mathbf{M}$ on the trajectory $\mathbf{s}_t$ can be expressed fully as $\mathbf{M} = \mathbf{M}(t) = \mathbf{M}(t; \mathbf{s}_{t_0})$, since $\mathbf{s}_{t_0}$ determines the trajectory $\mathbf{s}_t$.

Now let $\mathbf{M}_{t,t_0}$ denote the unique solution of Eq. (10) in $\mathbb{R}^{N \times N}$, over a finite time interval for which $\mathbf{s}_t$ exists in $\mathcal{S}$, that corresponds to the initial condition $\mathbf{M}_{t_0,t_0} = \mathbf{I}$. From the preceding discussion it follows that for each point $\mathbf{q}_{t_0} \in \mathbb{R}^N$, the linear equation

$$\frac{d\mathbf{q}}{dt} + \mathbf{F}(\mathbf{s}_t, t)\mathbf{q} = \mathbf{0} \tag{11}$$

has unique solution

$$\mathbf{q}_t = \mathbf{M}_{t,t_0}\mathbf{q}_{t_0} \in \mathbb{R}^N, \tag{12}$$

and that $\mathbf{q}_t$ has one continuous time derivative, for as long as $\mathbf{s}_t$ exists in $\mathcal{S}$. The linear equation (11) is called the (deterministic) perturbation equation associated with the original nonlinear dynamics (9). The matrix $\mathbf{M}_{t,t_0} = \mathbf{M}_{t,t_0}(\mathbf{s}_{t_0})$, which according to Eq. (12) expresses the solution of the perturbation equation directly in terms of its initial condition, is called the fundamental matrix, or solution operator, of the perturbation equation. The analogue of Eq. (11) for the stochastic initial value problem, which gives the evolution of stochastic initial perturbations under second-moment closure, is derived in Sec. 5 along with the corresponding mean and covariance evolution equations. Energetic consistency of the mean and covariance evolution equations is demonstrated in Sec. 6. The fundamental matrix of the stochastic perturbation equation has special properties due to this energetic consistency, which are described in Sec. 9.

It is well known (e.g. CODDINGTON and LEVINSON [1955, Ch. 1, Thm. 7.3]) that Eq. (10) can be solved explicitly for the determinant of $\mathbf{M}_{t,t_0}$:

$$\det \mathbf{M}_{t,t_0} = \exp\left[-\int_{t_0}^{t} \operatorname{tr} \mathbf{F}(\mathbf{s}_\tau, \tau)\, d\tau\right], \tag{13}$$

where $\operatorname{tr}\mathbf{F}$ denotes the trace, or sum of the diagonal elements, of $\mathbf{F}$. Define the symmetric and anti-symmetric (skew-symmetric) parts of $\mathbf{F}$ as $\mathbf{F}^s = \frac{1}{2}(\mathbf{F}+\mathbf{F}^T)$ and $\mathbf{F}^a = \frac{1}{2}(\mathbf{F}-\mathbf{F}^T)$, respectively, where the superscript T denotes transposition, so that $\mathbf{F} = \mathbf{F}^s + \mathbf{F}^a$ and $\mathbf{F}^T = \mathbf{F}^s - \mathbf{F}^a$. Since the diagonal elements of a real skew-symmetric matrix are all zero, $\mathbf{F}$ can be replaced in Eq. (13) by the symmetric matrix $\mathbf{F}^s$. This is one indication of the important role that the symmetric part of the Jacobian matrix plays in the dependence of trajectories on their initial points. In Sec. 6 it will be seen that the exchange of energy between the mean state and the stochastic perturbations occurs solely through $\mathbf{F}^s$. This role of $\mathbf{F}^s$ in the energetic coupling of the mean and covariance evolution equations is examined in detail in Secs. 7 and 8.

Another immediate consequence of Assumption 2 is that if $\partial\mathbf{f}/\partial t \in C(\mathcal{S}\times\mathcal{T})$, then in fact $\mathbf{s}_t$ has two continuous time derivatives, not just one. This follows by differentiating Eq. (9) once with respect to time and applying the chain rule:

$$\frac{d^2\mathbf{s}}{dt^2} - \mathbf{F}(\mathbf{s},t)\mathbf{f}(\mathbf{s},t) + \frac{\partial \mathbf{f}(\mathbf{s},t)}{\partial t} = \mathbf{0}.$$

Also, if $\partial\mathbf{f}/\partial t \in C(\mathcal{S}\times\mathcal{T})$, then Assumption 2 implies Assumption 1. It will not be assumed that $\partial\mathbf{f}/\partial t \in C(\mathcal{S}\times\mathcal{T})$, although this often does hold. For instance, in many problems $\mathbf{f}$ does not depend explicitly on time.

3 Conservation of Total Energy

Suppose now that one is given a nonlinear system

$$\frac{d\mathbf{x}}{dt} + \mathbf{g}(\mathbf{x}) = \mathbf{0}, \tag{14}$$

with state space $\mathcal{X}$, and with $\mathbf{g} : \mathcal{X} \times \mathcal{T} \to \mathbb{R}^N$ satisfying the hypotheses (Assumptions 1 and 2) imposed earlier on $\mathbf{f}$. Suppose also that there is a known function $\mathbf{s} : \mathcal{X} \to \mathbb{R}^N$ such that the "total energy"

$$E(t) = \tfrac{1}{2}\mathbf{s}^T(\mathbf{x}_t)\mathbf{s}(\mathbf{x}_t) \tag{15}$$

is conserved by the solutions of Eq. (14), that is,

$$\frac{dE}{dt} = \mathbf{s}^T(\mathbf{x}_t)\frac{d\mathbf{s}(\mathbf{x}_t)}{dt} = 0, \tag{16}$$

for each trajectory $\mathbf{x}_t$. Suppose finally that $\mathbf{s} = \mathbf{s}(\mathbf{x})$ defines a continuously differentiable coordinate transformation between the state space $\mathcal{X}$ and the range $\mathbf{s}(\mathcal{X})$, and denote by $\mathbf{A}(\mathbf{x}) = \partial\mathbf{s}/\partial\mathbf{x}$ the Jacobian matrix of this transformation.

Then for $\mathbf{s} = \mathbf{s}(\mathbf{x}_t)$ one has

$$\frac{d\mathbf{s}}{dt} = \mathbf{A}(\mathbf{x})\frac{d\mathbf{x}}{dt},$$

and so from Eq. (14) it follows that $\mathbf{s}(\mathbf{x}_t)$ satisfies the nonlinear system (9), for $\mathbf{f}$ given by

$$\mathbf{f}(\mathbf{s}, t) = \mathbf{A}\left(\mathbf{x}(\mathbf{s})\right)\mathbf{g}\left(\mathbf{x}(\mathbf{s}), t\right),$$

where $\mathbf{x} = \mathbf{x}(\mathbf{s})$ is the inverse transformation of $\mathbf{s} = \mathbf{s}(\mathbf{x})$. Furthermore, for Eq. (9), the total energy becomes simply $E = \frac{1}{2}\mathbf{s}^T\mathbf{s}$, and the statement (16) that the total energy is conserved becomes simply $\mathbf{s}^T\mathbf{f} = 0$. Thus, rather than considering Eq. (14) directly, for some general energy expression, in this article Eq. (9) is considered instead, under the simple hypothesis that the total energy $E = \frac{1}{2}\mathbf{s}^T\mathbf{s}$ is conserved:

Assumption 3 *For all* $\mathbf{s} \in \mathcal{S}$ *and* $t \in \mathcal{T}$*,*

$$\mathbf{s}^T\mathbf{f}(\mathbf{s}, t) = 0. \tag{17}$$

Transforming a general energy expression out of the problem in this way simplifies substantially the study of energetics of the second-moment closure equations. The cost is potentially a complicated expression for $\mathbf{f}$.

Assumption 3 says that the trajectories $\mathbf{s}_t$ of Eq. (9) satisfy

$$\mathbf{s}_t^T\mathbf{s}_t = \mathbf{s}_{t_0}^T\mathbf{s}_{t_0},$$

for as long as $t \in \mathcal{T}$ and the trajectory exists in $\mathcal{S}$. In geometrical terms, this means simply that each trajectory remains on the hyperspherical surface $\mathbf{s}^T\mathbf{s} = 2E$ in $\mathbb{R}^N$ on which it originates at time t_0. A trajectory can cease to exist only if this surface intersects the boundary of $\mathcal{S}$.

Since the change of coordinates to "energy variables" $\mathbf{s}$ is central to the results of this article, it is worthwhile to consider how it works in simple examples. Consider first the quadratically nonlinear system

$$\frac{du}{dt} + cu = 0,$$

$$\frac{d\phi}{dt} - 2cu^2 = 0,$$

with c a nonzero constant. Assumptions 1 and 2 are satisfied with the state space taken to be all of $\mathbb{R}^2$, and with arbitrarily large final time T for the interval $\mathcal{T} = [t_0, T]$. Therefore, for each initial condition in $\mathbb{R}^2$, this system has a unique solution in $\mathbb{R}^2$ which exists over arbitrarily long time intervals. In addition, for $E = \frac{1}{2}(u^2 + \phi)$ one has $dE/dt = 0$. This "energy" expression suggests that one consider also the state space $\mathcal{X}$ consisting of the upper half-plane $\phi > 0$ in (u, ϕ)-space, so that $E > 0$. The change of variables $s_1 = u$, $s_2 = \phi^{1/2}$ is a continuously differentiable coordinate transformation from $\mathcal{X}$ onto itself, and it yields Eq. (9) with

$$\mathbf{f} = c\,\frac{s_1}{s_2}\begin{bmatrix} s_2 \\ -s_1 \end{bmatrix},$$

which is singular along the s_1-axis $s_2 = 0$. Assumptions 1–3 are satisfied by this $\mathbf{f}$, with $\mathcal{S}$ being the upper half-plane $s_2 > 0$ in (s_1, s_2)-space, and again with arbitrarily large final time T for the interval $\mathcal{T}$. The "total energy" $E = \frac{1}{2}\mathbf{s}^T\mathbf{s}$ is conserved on each trajectory of Eq. (9), and every trajectory exists either until it hits the s_1-axis, which is the boundary of $\mathcal{S}$, or for all $t \in \mathcal{T}$ if it never hits the s_1-axis.

The solution of Eq. (9) in this example, for each $\mathbf{s}_{t_0} \in \mathcal{S}$, is just

$$s_1(t) = s_1(t_0)e^{-c(t-t_0)},$$

$$s_2(t) = \left[2E - s_1^2(t)\right]^{1/2}.$$

There is one stationary solution, which is $s_1 = 0$, $s_2 = (2E)^{1/2}$. If $c > 0$ and $s_1(t_0) \neq 0$, then $|s_1(t)|$ decreases monotonically toward zero, and $s_2(t)$ increases monotonically toward $(2E)^{1/2}$: all solutions approach the stationary one, and no trajectory can ever hit the s_1-axis. Thus if $c > 0$, then the trajectory corresponding to arbitrary $\mathbf{s}_{t_0} \in \mathcal{S}$ exists in $\mathcal{S}$ over arbitrarily long time intervals. If $c < 0$ on the other hand, and if $s_1(t_0) \neq 0$, then $|s_1(t)|$ increases monotonically toward $(2E)^{1/2}$ and $s_2(t)$ decreases monotonically toward zero, and the solution ceases to exist at time T_*,

$$T_* = t_0 - \frac{1}{2c}\log\left[2\left(1 + \frac{s_2^2(t_0)}{s_1^2(t_0)}\right)\right],$$

when $s_1^2(T_*) = 2E$, $s_2(T_*) = 0$, and $\mathbf{f}$ is no longer defined. However, for the problem in the original variables u and ϕ, as posed on all of $\mathbb{R}^2$, nothing bad can happen at this time T_*, nor at any other finite time, since it was already seen that every trajectory of the original problem exists in $\mathbb{R}^2$ over arbitrarily long time intervals. In fact, all that happens is that the trajectory continues downward into the lower half-plane $\phi < 0$ in (u, ϕ)-space, along the parabola $\phi_t = 2E - u_t^2$. Thus it is possible for long-time existence of solutions to be lost through the transformation to energy variables.

Fortunately, for spatially discretized versions of the main first-order hyperbolic partial differential equations of atmospheric and ocean dynamics, nothing need be lost in the transformation to energy variables, since all that is required in the transformation is to take the square root of mass (gravitational potential energy) and temperature (internal energy) variables that are required on physical grounds to remain positive. As a simple example, consider the shallow-water equations in one space dimension, taken to be periodic. These are usually written as the momentum equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + \frac{\partial \phi}{\partial x} = 0,$$

and the continuity equation

$$\frac{\partial \phi}{\partial t} + \frac{\partial u\phi}{\partial x} = 0,$$

where u is the speed, ϕ is the geopotential, and $x \in [0, L]$ is the space variable. From the differential equations it follows that solutions satisfy the energy equation

$$\frac{\partial(\phi u^2 + \phi^2)}{\partial t} + \frac{\partial(\phi u^3 + 2\phi^2 u)}{\partial x} = 0,$$

which on the periodic domain $[0, L]$ implies conservation of the total energy

$$E = \tfrac{1}{2}\int_0^L \left(\phi u^2 + \phi^2\right)\, dx.$$

Reasonable spatial discretization of the variables u and ϕ gives rise to a system of ordinary differential equations (14) that conserves a discretized version (15) of this energy integral.

The change of variable from u to $\alpha = \phi^{1/2}u$ transforms the momentum equation to

$$\frac{\partial \alpha}{\partial t} + u\frac{\partial \alpha}{\partial x} + \tfrac{1}{2}\frac{\partial u}{\partial x}\alpha + \phi^{1/2}\frac{\partial \phi}{\partial x} = 0,$$

and substituting $u = \phi^{-1/2}\alpha$ here and in the continuity equation gives the shallow-water system in terms of the variables α and ϕ alone. Reasonable spatial discretization of the transformed system, say with $\alpha_n(t) = \alpha(x_n, t)$, $\phi_n(t) = \phi(x_n, t)$, and $x_n = nL/N$, for $n = 1, \ldots, N$, gives a system (9) for $\mathbf{s} = [\alpha_1, \ldots, \alpha_N, \phi_1, \ldots, \phi_N]^T$ that conserves the discretized total energy $E = \frac{1}{2}\mathbf{s}^T\mathbf{s}\frac{L}{N}$. One can see that, because of the substitution $u = \phi^{-1/2}\alpha$, the function $\mathbf{f}$ will be singular on each of the hyperplanes $\phi_n = 0$, $n = 1, \ldots, N$, in $\mathbb{R}^{2N}$. Merely taking $\mathcal{S}$ to be the convex open set

$$\mathcal{S} = \{\mathbf{s} = [\alpha_1, \ldots, \alpha_N, \phi_1, \ldots, \phi_N]^T : \phi_n > 0, n = 1, \ldots, N\}$$

does not, however, give rise to an initial-value problem whose trajectories exist in this $\mathcal{S}$ over arbitrarily long time intervals. That is, trajectories starting in

this $\mathcal{S}$ may hit the boundary of $\mathcal{S}$, developing zero geopotential somewhere, at some time $T_* < T$, and thereby cease to exist.

For the one-dimensional shallow-water equations, it is straightforward to define a state space such that every trajectory does exist over arbitrarily long time intervals, for many spatial discretizations of either the (u, ϕ) system or the (α, ϕ) system. It follows from the shallow-water equations that the characteristic speeds

$$c_+ = u + \phi^{1/2} = \phi^{-1/2}(\alpha + \phi),$$
$$c_- = u - \phi^{1/2} = \phi^{-1/2}(\alpha - \phi),$$

satisfy the coupled advection equations

$$\frac{\partial c_+}{\partial t} + (\tfrac{3}{4}c_+ + \tfrac{1}{4}c_-)\frac{\partial c_+}{\partial x} = 0,$$
$$\frac{\partial c_-}{\partial t} + (\tfrac{1}{4}c_+ + \tfrac{3}{4}c_-)\frac{\partial c_-}{\partial x} = 0.$$

Therefore if initially $c_+ > 0$ and $c_- < 0$, equivalently $\phi > |\alpha|$, for all $x \in [0, L]$, then this remains true for all time. Positivity-preserving (of c_+ and $-c_-$) discretizations (e.g. LIN et al. [1994]) of the shallow-water equations therefore have the property that trajectories with initial points in the convex open set

$$\mathcal{S}_0 = \{\mathbf{s} = [\alpha_1, \ldots, \alpha_N, \phi_1, \ldots, \phi_N]^T : \phi_n > |\alpha_n|, n = 1, \ldots, N\}$$

exist in $\mathcal{S}_0$ over arbitrarily long time intervals. More generally, if there is a constant $c > 0$ such that initially $c_+ > c$ and $c_- < -c$ (equivalently $\phi^{1/2} > c + |u|$, or $\phi^{1/2} > c/2 + (c^2/4 + |\alpha|)^{1/2}$) for all $x \in [0, L]$, the coupled advection equations imply that this also remains true for all time. Therefore, for appropriate spatial discretizations, trajectories with initial points in the convex open set

$$\mathcal{S}_c = \{\mathbf{s} = [\alpha_1, \ldots, \alpha_N, \phi_1, \ldots, \phi_N]^T : \phi_n^{1/2} > c/2 + (c^2/4 + |\alpha_n|)^{1/2}, n = 1, \ldots, N\}$$

exist in $\mathcal{S}_c$ over arbitrarily long time intervals. On the state space $\mathcal{S}_0$, the geopotential can approach zero arbitrarily closely wherever $u = 0$. On the state space $\mathcal{S}_c$ with $c > 0$, however, the geopotential is guaranteed to be strictly positive everywhere, $\phi > (c + |u|)^2 \geq c^2$, for all $t \in \mathcal{T}$, and therefore the total energy is also strictly positive,

$$E = \tfrac{1}{2}\int_0^L \left(\alpha^2 + \phi^2\right)\, dx \geq \tfrac{1}{2}\int_0^L \phi^2\, dx > \tfrac{1}{2}Lc^4. \tag{18}$$

4 The Stochastic Initial-Value Problem

Now let Ω be the sample space of a complete probability space, and denote by $\mathcal{E}$ the expectation operator on the probability space. A random vector $\mathbf{s} \in \mathbb{R}^N$ is a vector function $\mathbf{s} = \mathbf{s}(\omega)$ of the probability variable $\omega \in \Omega$, that is, $\mathbf{s} : \Omega \to \mathbb{R}^N$. For fixed ω, $\mathbf{s}(\omega)$ is called a realization of the random vector $\mathbf{s}$.

Consider a second-order random vector, that is, a random vector $\mathbf{s} \in \mathbb{R}^N$ such that $\mathcal{E}\mathbf{s}^T\mathbf{s} < \infty$. By the Schwarz inequality one has

$$(\mathcal{E}\mathbf{s})^T(\mathcal{E}\mathbf{s}) \leq (\mathcal{E}|\mathbf{s}|)^T(\mathcal{E}|\mathbf{s}|) \leq \mathcal{E}\mathbf{s}^T\mathbf{s},$$

and therefore the mean $\overline{\mathbf{s}} = \mathcal{E}\mathbf{s}$ exists and is finite. Denote the departure from the mean by $\mathbf{s}' = \mathbf{s} - \overline{\mathbf{s}}$. Since $\mathcal{E}\mathbf{s}' = \mathbf{0}$ one has

$$\mathcal{E}\mathbf{s}^T\mathbf{s} = \overline{\mathbf{s}}^T\overline{\mathbf{s}} + \mathcal{E}\mathbf{s}'^T\mathbf{s}',$$

and therefore the total variance V defined by

$$V = \mathcal{E}\mathbf{s}'^T\mathbf{s}'$$

is also finite, $V \leq \mathcal{E}\mathbf{s}^T\mathbf{s}$. Since the total variance is finite, it follows by a further application of the Schwarz inequality that the elements of the $N \times N$ covariance matrix $\mathbf{P}$ defined by

$$\mathbf{P} = \mathcal{E}\mathbf{s}'\mathbf{s}'^T$$

also exist and are finite. The covariance matrix is, by definition, symmetric and positive semidefinite. Also, the total variance is just the trace of the covariance matrix, $V = \operatorname{tr}\mathbf{P}$. Associating randomness with uncertainty, one can say that the total variance of a second-order random vector is a scalar measure of its uncertainty.

For a random vector $\mathbf{s} \in \mathbb{R}^N$, one writes $\mathbf{s} \in \mathcal{S}$ wp1 (with probability one) if $\mathbf{s}(\omega) \in \mathcal{S}$ for all $\omega \in \Omega$, except possibly for an ω set of probability measure zero. A random vector $\mathbf{s} \in \mathcal{S}$ wp1 will be called a random state. Every second-order random state $\mathbf{s}$ has mean $\overline{\mathbf{s}} \in \mathcal{S}$, because $\mathcal{S}$ was taken to be convex.

In the stochastic initial-value problem treated here, Eq. (9) is considered for second-order random initial states $\mathbf{s}_{t_0}$, that is, random initial states with

$$\mathcal{E}E_{t_0} = \tfrac{1}{2}\mathcal{E}\mathbf{s}_{t_0}^T\mathbf{s}_{t_0} < \infty.$$

Denote by $\mathbf{s}_t(\omega)$ the trajectory corresponding to a realization $\mathbf{s}_{t_0}(\omega) \in \mathcal{S}$. From Assumption 3 it follows that $\mathbf{s}_t(\omega)$ satisfies the total energy conservation equation

$$\mathbf{s}_t^T(\omega)\mathbf{s}_t(\omega) = \mathbf{s}_{t_0}^T(\omega)\mathbf{s}_{t_0}(\omega), \tag{19}$$

for each $\omega \in \Omega$ such that $\mathbf{s}_{t_0}(\omega) \in \mathcal{S}$. If the deterministic initial-value problem is such that all trajectories starting in $\mathcal{S}$ exist in $\mathcal{S}$ over arbitrarily long time intervals $\mathcal{T} = [t_0, T]$, for instance if $\mathcal{S} = \mathbb{R}^N$, then the realizations $\mathbf{s}_t(\omega)$ with $\mathbf{s}_{t_0}(\omega) \in \mathcal{S}$ define a random vector $\mathbf{s}_t \in \mathcal{S}$ wp1, for all $t \in \mathcal{T}$, and by taking expectations in Eq. (19) it follows that $\mathbf{s}_t$ is also a second-order random vector, for all $t \in \mathcal{T}$. In this case, therefore, $\mathbf{s}_t$ has finite mean $\overline{\mathbf{s}}_t \in \mathcal{S}$ and covariance matrix $\mathbf{P}_t \in \mathbb{R}^{N\times N}$, for all $t \in \mathcal{T}$. The family of random vectors $\{\mathbf{s}_t : t \in \mathcal{T}\}$ is called a (second-order) stochastic process. The trajectories $\mathbf{s}_t(\omega)$, for fixed $\omega \in \Omega$, are called the sample functions, or sample paths, of the process. By Assumption 1, each sample path with $\mathbf{s}_{t_0}(\omega) \in \mathcal{S}$ is in $C^1(\mathcal{T})$. In other words, the sample paths are in $C^1(\mathcal{T})$ wp1.

In case the deterministic initial-value problem is such that there are trajectories starting in $\mathcal{S}$ that hit the boundary of $\mathcal{S}$ before time T, one can restrict the problem to initial states in some prescribed subset $\mathcal{S}' \subset \mathcal{S}$, not necessarily convex or open, but whose boundary nowhere touches the boundary of $\mathcal{S}$. Since the trajectories are continuous paths in state space, traversed at finite speed, there is then a time $T' > t_0$ such that the trajectory $\mathbf{s}_t(\omega)$ corresponding to an arbitrary point $\mathbf{s}_{t_0}(\omega) \in \mathcal{S}'$ exists in $\mathcal{S}$ for all time t in the closed interval $\mathcal{T}' = [t_0, T']$, with T' independent of $\mathbf{s}_{t_0}(\omega)$. Thus in this case, the stochastic initial-value problem is restricted to second-order random initial vectors $\mathbf{s}_{t_0} \in \mathcal{S}'$ wp1. The realizations $\mathbf{s}_t(\omega)$ with $\mathbf{s}_{t_0}(\omega) \in \mathcal{S}'$ still satisfy Eq. (19), but now only for $t \in \mathcal{T}'$. Thus they define a second-order stochastic process $\{\mathbf{s}_t : t \in \mathcal{T}'\}$, with $\mathbf{s}_t \in \mathcal{S}$ wp1, and with finite mean $\overline{\mathbf{s}}_t \in \mathcal{S}$ and covariance matrix $\mathbf{P}_t \in \mathbb{R}^{N\times N}$, for all $t \in \mathcal{T}'$. The sample paths are in $C^1(\mathcal{T}')$ wp1.

Thus two cases have been distinguished for the stochastic initial-value problem. In the first one, Eq. (9) defines a second-order stochastic process $\{\mathbf{s}_t : t \in \mathcal{T}\}$ for each second-order random initial state. In the second case, Eq. (9) defines a second-order stochastic process $\{\mathbf{s}_t : t \in \mathcal{T}'\}$ for each second-order random initial vector $\mathbf{s}_{t_0}$ such that $\mathbf{s}_{t_0} \in \mathcal{S}'$ wp1. Denote by V_t the total variance

$$V_t = \mathcal{E}\mathbf{s}_t'^T\mathbf{s}_t' = \operatorname{tr}\mathbf{P}_t,$$

for $t \in \mathcal{T}$ in the first case, and for $t \in \mathcal{T}'$ in the second. Taking expectations in Eq. (19) gives

$$\overline{\mathbf{s}}_t^T\overline{\mathbf{s}}_t + V_t = \overline{\mathbf{s}}_{t_0}^T\overline{\mathbf{s}}_{t_0} + V_{t_0}, \tag{20}$$

for $t \in \mathcal{T}$ in the first case, and for $t \in \mathcal{T}'$ in the second. The total variance V_t is a scalar measure of the uncertainty present in the random state $\mathbf{s}_t$ due to uncertainty in the random initial state $\mathbf{s}_{t_0}$.

Equation (20) says that the uncertainty in solutions of Eq. (9) due to uncertainty in the initial condition, as measured by the total variance, can increase (decrease) only as a result of extracting energy from (inserting energy into) the mean state $\overline{\mathbf{s}}$, with the change in total variance balanced exactly by twice the change in total energy $\frac{1}{2}\overline{\mathbf{s}}^T\overline{\mathbf{s}}$ of the mean state. This is purely a consequence of conservation of total energy for the nonlinear dynamics, and it holds regardless of any assumptions one might make on the form of the probability distribution of the random initial state $\mathbf{s}_{t_0}$, apart from existence of its first two moments. In fact, Eq. (20) holds even if no moments beyond the first two exist at any time. It is simply a statement about second-order stochastic processes whose realizations satisfy Eq. (19) with probability one.

Equation (20) implies in particular the bound

$$V_t \le \overline{\mathbf{s}}_{t_0}^T\overline{\mathbf{s}}_{t_0} + V_{t_0}, \tag{21}$$

valid for all time $t \in \mathcal{T}$ in the first case, $t \in \mathcal{T}'$ in the second, with equality holding at some particular time τ if, and only if, all the energy of the mean state has been extracted at that time, $\overline{\mathbf{s}}_\tau = \mathbf{0}$. This is a simple, general statement of the maximum level of uncertainty that can occur in solutions of Eq. (9).

There is no implication, however, that the uncertainty actually saturates, that is, becomes or approaches a constant in time, either at the level $\overline{\mathbf{s}}_{t_0}^T \overline{\mathbf{s}}_{t_0} + V_{t_0}$ or at any other level. Saturation of uncertainty will be discussed briefly at the end of Sec. 8.

For many dynamical systems, the state space is naturally bounded away from the origin of coordinates $\mathbf{s} = \mathbf{0}$ in $\mathbb{R}^N$, so that in fact the mean state cannot vanish at any time and the bound (21) can be improved upon. This is the case when there are mass-like and/or temperature-like state variables, for instance, that are constrained by the physics of the problem to be bounded from below by positive constants. If every point $\mathbf{s}$ in the state space $\mathcal{S}$ of the problem satisfies an inequality $\mathbf{s}^T\mathbf{s} > 2E_{\min} \geq 0$, then since $\overline{\mathbf{s}}_t \in \mathcal{S}$ one has $\overline{\mathbf{s}}_t^T \overline{\mathbf{s}}_t > 2E_{\min} \geq 0$; here $2E_{\min}$ is just the minimum Euclidean distance from the origin to the boundary of $\mathcal{S}$. In this case Eq. (20) implies the stronger bound

$$V_t < \overline{\mathbf{s}}_{t_0}^T \overline{\mathbf{s}}_{t_0} + V_{t_0} - 2E_{\min}. \tag{22}$$

For example, in Sec. 3 it was shown that for appropriately discretized versions of the one-dimensional shallow-water equations with state space $\mathcal{S}_c$, $c \geq 0$, the geopotential satisfies $\phi > c^2$ for all $t \in \mathcal{T}$, and since the discretized total energy was $E = \frac{1}{2}\mathbf{s}^T\mathbf{s}\frac{L}{N}$, Eq. (18) gives $\overline{\mathbf{s}}_t^T \overline{\mathbf{s}}_t > Nc^4$. Thus for appropriate discrete shallow-water dynamics on $\mathcal{S}_c$ one has

$$V_t < \overline{\mathbf{s}}_{t_0}^T \overline{\mathbf{s}}_{t_0} + V_{t_0} - Nc^4, \tag{23}$$

for all $t \in \mathcal{T}$.

Since Eq. (20) holds without any assumptions on moments beyond the first two, even without their existence, one expects to be able to find approximate evolution equations for just the mean and covariance matrix, such that their solutions are guaranteed to satisfy Eq. (20) and therefore also the bounds (21) and (22). A closed system of differential equations for the mean and covariance matrix that has this property will be said to be *energetically consistent*. The second-moment closure equations derived in Sec. 5 constitute a closed, nonlinearly coupled system of differential equations for the mean and covariance matrix. In Sec. 6 it will be shown that the nonlinear coupling term in the second-moment closure equation for the mean makes them energetically consistent. First the exact equations for the mean and covariance matrix, which are not closed unless the dynamics are linear, will be derived.

If E represents a physical total energy, then for both the deterministic and stochastic initial-value problems there seems little reason to consider initial points in $\mathbb{R}^N$ with total energy greater than or equal to some prescribed, fixed amount, say $E \geq E_{\max}$. Such points will be eliminated from consideration by making the following simple hypothesis on $\mathcal{S}$:

Assumption 4 *$\mathcal{S} \subseteq \mathcal{S}_{\max}$, where $\mathcal{S}_{\max}$ denotes the interior of the hypersphere $\mathbf{s}^T\mathbf{s} = 2E_{\max}$ in $\mathbb{R}^N$:*

$$\mathcal{S}_{\max} = \{\mathbf{s} \in \mathbb{R}^N : \mathbf{s}^T\mathbf{s} < 2E_{\max}\}.$$

This assumption imposes no restriction on existence of solutions, since no trajectory starting in $\mathcal{S} \subseteq \mathcal{S}_{\max}$ can ever hit the boundary of $\mathcal{S}_{\max}$ at any time, by conservation of total energy. It is, however, a restriction on the initial probability distribution, and therefore on the probability distribution at any time, because it implies that if $\mathbf{s}$ is a random state, then $\mathbf{s}^T\mathbf{s} < 2E_{\max}$ wp1. In particular, no random state is (multivariate) normally distributed, and the marginal distributions of a random state also cannot be normal. This may seem a significant restriction. However, it is suggested by the physical problem at hand. Moreover, as discussed already, for atmospheric and ocean dynamics there are also usually state variables that are constrained to be positive, in fact often bounded from below by positive constants, and these cannot be normally distributed.

An immediate consequence of Assumption 4 is that every random state $\mathbf{s}$ is a second-order random vector, in fact, $\mathcal{E}\mathbf{s}^T\mathbf{s} < 2E_{\max}$. Also, since $\mathbf{f} \in C(\mathcal{S}\times\mathcal{T})$ by Assumption 1, it follows from Assumption 4 that $\mathbf{f}^T\mathbf{f} < \infty$ on $\mathcal{S}\times\mathcal{T}$. Therefore, if $\mathbf{s}$ is a random state, then $\mathbf{f}(\mathbf{s})$ is a second-order random vector, $\mathcal{E}\mathbf{f}^T(\mathbf{s})\mathbf{f}(\mathbf{s}) < \infty$.

Again let $\mathbf{s}_t$ (either for $t \in \mathcal{T}$ or $t \in \mathcal{T}'$, depending on the case) denote the solution of the stochastic initial-value problem for Eq. (9). Then $\mathcal{E}|\mathbf{f}(\mathbf{s}_t, t)| < \infty$ since $\mathbf{f}(\mathbf{s}_t, t)$ is a second-order random vector, and therefore

$$\int_{t_0}^{t} \mathcal{E}|\mathbf{f}(\mathbf{s}_\tau, \tau)|\, d\tau < \infty.$$

It follows (e.g. DOOB [1953, Thm. 2.7, p. 62]) that

$$\mathcal{E}\int_{t_0}^{t} \mathbf{f}(\mathbf{s}_\tau, \tau)\, d\tau = \int_{t_0}^{t} \mathcal{E}\mathbf{f}(\mathbf{s}_\tau, \tau)\, d\tau, \tag{24}$$

where both integrals exist and are finite. Now write Eq. (9) as the integral equation

$$\mathbf{s}_t(\omega) = \mathbf{s}_{t_0}(\omega) - \int_{t_0}^{t} \mathbf{f}(\mathbf{s}_\tau(\omega), \tau)\, d\tau. \tag{25}$$

Taking expectations and using Eq. (24) gives

$$\overline{\mathbf{s}}_t = \overline{\mathbf{s}}_{t_0} - \int_{t_0}^{t} \overline{\mathbf{f}}(\mathbf{s}_\tau, \tau)\, d\tau, \tag{26}$$

where $\overline{\mathbf{f}}(\mathbf{s}_\tau, \tau) = \mathcal{E}\mathbf{f}(\mathbf{s}_\tau, \tau)$, which shows that the mean state $\overline{\mathbf{s}}_t$ is a continuous function of time. Differentiating in Eq. (26) gives an equation for the mean state in differential form,

$$\frac{d\overline{\mathbf{s}}_t}{dt} + \overline{\mathbf{f}}(\mathbf{s}_t, t) = \mathbf{0}, \tag{27}$$

which is satisfied almost everywhere in $\mathcal{T}$ in the first case, that is, except possibly on a subset of $\mathcal{T}$ of Lebesgue measure zero, and almost everywhere in $\mathcal{T}'$ in the second. This exact equation for the mean state is not a differential equation unless $\mathbf{f}(\mathbf{s}, t)$ is linear in $\mathbf{s}$, in which case $\overline{\mathbf{f}}(\mathbf{s}_t, t) = \mathbf{f}(\overline{\mathbf{s}}_t, t)$. Also, $\overline{\mathbf{f}}(\mathbf{s}_t, t)$ is

not necessarily continuous in time, and so $\overline{\mathbf{s}}_t$ is not necessarily continuously differentiable. Comparing Eqs. (9) and (27) shows that the commutation

$$\mathcal{E}\frac{d\mathbf{s}_t}{dt} = \frac{d\overline{\mathbf{s}}_t}{dt} \tag{28}$$

holds, almost everywhere in time.

To derive the equation for the covariance matrix, first subtract Eq. (26) from Eq. (25), to obtain

$$\mathbf{s}'_t(\omega) = \mathbf{s}'_{t_0}(\omega) - \int_{t_0}^{t} \mathbf{f}'(\mathbf{s}_\tau(\omega), \tau)\, d\tau, \tag{29}$$

where $\mathbf{s}' = \mathbf{s} - \overline{\mathbf{s}}$ and $\mathbf{f}' = \mathbf{f} - \overline{\mathbf{f}}$. Postmultiplying this equation by the transpose of itself, then taking expectations, and then exchanging the order of expectation and integration, which again can be justified by Assumption 4, gives

$$\begin{aligned}\mathbf{P}_t = \mathbf{P}_{t_0} \quad &- \quad \int_{t_0}^{t} \mathcal{E}\left[\mathbf{f}'(\mathbf{s}_\tau, \tau)\mathbf{s}'^T_{t_0}\right]\, d\tau - \int_{t_0}^{t} \mathcal{E}\left[\mathbf{s}'_{t_0}\mathbf{f}'^T(\mathbf{s}_\tau, \tau)\right]\, d\tau \\ &+ \quad \int_{t_0}^{t}\int_{t_0}^{t} \mathcal{E}\left[\mathbf{f}'(\mathbf{s}_{\tau_1}, \tau_1)\mathbf{f}'^T(\mathbf{s}_{\tau_2}, \tau_2)\right]\, d\tau_1\, d\tau_2,\end{aligned}$$

which shows that the covariance matrix $\mathbf{P}_t$ is a continuous function of time. Differentiating this result and using Eq. (29) gives an equation for the covariance matrix in differential form,

$$\frac{d\mathbf{P}_t}{dt} + \mathcal{E}\left[\mathbf{f}'(\mathbf{s}_t, t)\mathbf{s}'^T_t\right] + \mathcal{E}\left[\mathbf{f}'(\mathbf{s}_t, t)\mathbf{s}'^T_t\right]^T = 0, \tag{30}$$

satisfied almost everywhere in $\mathcal{T}$, or in $\mathcal{T}'$. Again, this is not a differential equation unless $\mathbf{f}$ is linear, and the elements of $\mathbf{P}_t$ are not necessarily continuously differentiable. Equation (30) can be used to show that the commutation

$$\mathcal{E}\frac{d\mathbf{s}'_t\mathbf{s}'^T_t}{dt} = \frac{d\mathbf{P}_t}{dt} \tag{31}$$

holds, almost everywhere in time.

Equations (27) and (30) are usually derived under an hypothesis much weaker than Assumption 4 which, like Assumption 4, also implies that $\mathbf{f}(\mathbf{s}_t, t)$ is a second-order random vector (e.g. DOOB [1953, p. 277, hypothesis H_2], JAZWINSKI [1970, p. 105, hypothesis H_1]). However, Assumption 4 makes sense for conservative dynamics, and it greatly simplifies the derivation of Eqs. (27) and (30). It also makes for little difference between the formulation of the stochastic initial-value problem and that of the deterministic initial-value problem, since it makes every random state a second-order random vector. All that has been necessary for the stochastic problem was to restrict initial random vectors to lie in some set $\mathcal{S}'$ (wp1) contained wholly in the interior of $\mathcal{S}$ in case solutions of the deterministic initial-value problem do not exist over arbitrarily long time intervals, to ensure that all the sample paths exist for some minimum amount of time $T' - t_0 > 0$ wp1.

5 The Second-Moment Closure Equations

Two final hypotheses on $\mathbf{f}$ will now be made:

Assumption 5 *$\partial^2\mathbf{f}(\mathbf{s},t)/\partial s_j \partial s_k \in C(\mathcal{S}\times\mathcal{T})$, for $j,k=1,\ldots,N$.*

Assumption 6 *The second partial derivatives of $\mathbf{f}$ are Lipschitz continuous in $\mathbf{s}$ on $\mathcal{S}\times\mathcal{T}$, uniformly in t. That is, there are constants K_{jk} such that*

$$\left\| \frac{\partial^2\mathbf{f}(\mathbf{s}_1,t)}{\partial s_j \partial s_k} - \frac{\partial^2\mathbf{f}(\mathbf{s}_2,t)}{\partial s_j \partial s_k} \right\| \leq K_{jk} \left\| \mathbf{s}_1 - \mathbf{s}_2 \right\|,$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^N$, for each $\mathbf{s}_1,\mathbf{s}_2 \in \mathcal{S}$, for each $t\in\mathcal{T}$, and for $j,k=1,\ldots,N$.

Let $\mathbf{s}$ be a random state, that is, $\mathbf{s}\in\mathcal{S}$ wp1. Since $\overline{\mathbf{s}} = \mathcal{E}\mathbf{s} \in \mathcal{S}$, it follows from Assumption 5 that $\mathbf{f}(\mathbf{s}) = \mathbf{f}(\mathbf{s},t)$ has a Taylor expansion about $\overline{\mathbf{s}}$ up to second order,

$$\mathbf{f}(\mathbf{s}) = \mathbf{f}(\overline{\mathbf{s}}) + \mathbf{F}(\overline{\mathbf{s}})\mathbf{s}' + \tfrac{1}{2}\sum_j\sum_k \frac{\partial^2\mathbf{f}(\overline{\mathbf{s}})}{\partial s_j \partial s_k} s_j' s_k' + \mathbf{g}(\mathbf{s}',\overline{\mathbf{s}}), \tag{32}$$

where $\mathbf{F} = \mathbf{F}(\mathbf{s}) = \mathbf{F}(\mathbf{s},t)$ is the Jacobian matrix of $\mathbf{f}$ introduced in Sec. 2, and where s_j' is the j^{th} element of the random vector $\mathbf{s}' = \mathbf{s} - \overline{\mathbf{s}}$. It follows from Assumption 6 that the remainder term $\mathbf{g}$ is $\mathcal{O}((\mathbf{s}')^3)$ for each fixed $\overline{\mathbf{s}}\in\mathcal{S}$. It followed from Assumption 4 that $\mathbf{s}$ is a second-order random vector, and along with Assumption 1 that $\mathbf{f}(\mathbf{s})$ is also a second-order random vector, so that $\overline{\mathbf{f}} = \mathcal{E}\mathbf{f}$ exists and is finite. Taking expectations in the Taylor expansion then shows that $\overline{\mathbf{g}} = \mathcal{E}\mathbf{g}$ also exists and is finite, and that

$$\overline{\mathbf{f}} = \mathbf{f}(\overline{\mathbf{s}}) + \tfrac{1}{2}\sum_j\sum_k \frac{\partial^2\mathbf{f}(\overline{\mathbf{s}})}{\partial s_j \partial s_k} P_{jk} + \overline{\mathbf{g}}, \tag{33}$$

where $P_{jk} = \mathcal{E}s_j' s_k'$ is the $(j,k)^{th}$ element of the covariance matrix $\mathbf{P} = \mathcal{E}\mathbf{s}'\mathbf{s}'^T$.

The *mean equation* is obtained by substituting Eq. (33) into Eq. (27) and neglecting the remainder term $\overline{\mathbf{g}}$, to yield

$$\frac{d\overline{\mathbf{s}}}{dt} + \mathbf{f}(\overline{\mathbf{s}}) + \tfrac{1}{2}\sum_j\sum_k \frac{\partial^2\mathbf{f}(\overline{\mathbf{s}})}{\partial s_j \partial s_k} P_{jk} = \mathbf{0}, \tag{34}$$

where the time subscripts have been omitted because the mean equation is only an approximate equation for the evolution of $\overline{\mathbf{s}}_t$. The *covariance evolution equation* is obtained by using Eqs. (32) and (33) to approximate $\mathbf{f}' = \mathbf{f}'(\mathbf{s}) = \mathbf{f}(\mathbf{s}) - \overline{\mathbf{f}}$ by $\mathbf{F}(\overline{\mathbf{s}})\mathbf{s}'$, and substituting this into Eq. (30), yielding

$$\frac{d\mathbf{P}}{dt} + \mathbf{F}(\overline{\mathbf{s}})\mathbf{P} + \mathbf{P}\mathbf{F}^T(\overline{\mathbf{s}}) = \mathbf{0}, \tag{35}$$

where the time subscripts have again been omitted because the covariance evolution equation is only an approximate equation for the evolution of $\mathbf{P}_t$.

Equations (34) and (35) together constitute the *second-moment closure equations* for the evolution of the mean state and covariance matrix. In case $\mathbf{f} = \mathbf{f}(\mathbf{s}, t)$ is linear in its first argument, the second-moment closure equations decouple and they are exact: the Jacobian matrix $\mathbf{F} = \mathbf{F}(\mathbf{s}, t)$ is independent of its first argument, so that the covariance evolution equation decouples from the mean equation, and the second partial derivatives of $\mathbf{f}(\mathbf{s}, t)$ with respect to their first argument all vanish, so that the mean equation likewise decouples from the covariance evolution equation. In case $\mathbf{f}$ is nonlinear, the mean and covariance evolution equations are fully coupled, in both directions.

Let $\mathcal{P} \subseteq \mathbb{R}^{N \times N}$ be an open set, and suppose that $\overline{\mathbf{s}}_{t_0} \in \mathcal{S}$ and $\mathbf{P}_{t_0} \in \mathcal{P}$. Then there is a half-open time interval $\mathcal{T}_0 = [t_0, T_0)$, with T_0 depending in general on $\overline{\mathbf{s}}_{t_0}$ and $\mathbf{P}_{t_0}$, such that there exists a unique solution $(\overline{\mathbf{s}}(t) \in \mathcal{S}, \mathbf{P}(t) \in \mathcal{P})$ of this coupled system for all $t \in \mathcal{T}_0$, satisfying initial condition $(\overline{\mathbf{s}}(t_0), \mathbf{P}(t_0)) = (\overline{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0})$. Assumptions 1, 2, and 5 guarantee that there exists at least one solution on $\mathcal{T}_0$ satisfying the initial condition, and also that every solution on $\mathcal{T}_0$ is in $C^1(\mathcal{T}_0)$. Assumptions 2 and 6 guarantee that there exists at most one solution on $\mathcal{T}_0$ satisfying the initial condition. As in Sec. 4, if $\overline{\mathbf{s}}_{t_0}$ and $\mathbf{P}_{t_0}$ are restricted to lie in some subsets $\overline{\mathcal{S}} \subset \mathcal{S}$ and $\overline{\mathcal{P}} \subset \mathcal{P}$, respectively, whose boundaries nowhere touch the boundaries of $\mathcal{S}$ and $\mathcal{P}$, then existence and uniqueness of solutions are assured over some closed time interval $\overline{\mathcal{T}} = [t_0, \overline{T}]$ with $\overline{T} > t_0$. Thus if one is interested in solving Eqs. (34) and (35) with some particular $\overline{\mathbf{s}}_{t_0} \in \mathcal{S}$ given, as is often the case, then existence and uniqueness on some closed time interval $\overline{\mathcal{T}} = [t_0, \overline{T}]$ are assured for any $\mathbf{P}_{t_0} \in \overline{\mathcal{P}}$, simply by taking $\overline{\mathcal{S}} = \{\overline{\mathbf{s}}_{t_0}\}$.

In fact, Eqs. (34) and (35) are supposed to be solved for initial condition $(\overline{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0})$ being the mean state and covariance matrix of a random state $\mathbf{s}_{t_0}$, if the equations are to approximate the evolution of the mean and covariance matrix of the random state $\mathbf{s}_t$. In particular, $\mathbf{P}_{t_0}$ is supposed to be symmetric positive semidefinite. Moreover, this means that $\mathbf{P}_{t_0}$ cannot be specified independently of $\overline{\mathbf{s}}_{t_0}$. For instance, under Assumption 4 one has the simple restriction

$$\overline{\mathbf{s}}_{t_0}^T \overline{\mathbf{s}}_{t_0} + \operatorname{tr} \mathbf{P}_{t_0} < 2E_{\max}. \tag{36}$$

The particular geometry of $\mathcal{S}$ for a given problem can restrict $\mathbf{P}_{t_0}$ in terms of $\overline{\mathbf{s}}_{t_0}$ much more severely than this, with the restriction being the strongest when $\overline{\mathbf{s}}_{t_0}$ is near the boundary of $\mathcal{S}$, as discussed further below. Since $\mathbf{P}_{t_0}$ is not supposed to be specified independently of $\overline{\mathbf{s}}_{t_0}$, it is simplest to pose the initial-value problem for a given, fixed $\overline{\mathbf{s}}_{t_0} \in \mathcal{S}$, then to define the set $\mathcal{P} = \mathcal{P}(\overline{\mathbf{s}}_{t_0})$ of symmetric positive semidefinite matrices over which $\mathbf{P}_{t_0}$ is allowed to vary due to the particular geometry of $\mathcal{S}$, and finally to define a subset $\overline{\mathcal{P}} = \overline{\mathcal{P}}(\overline{\mathbf{s}}_{t_0}) \subset \mathcal{P}(\overline{\mathbf{s}}_{t_0})$ whose boundary nowhere touches that of $\mathcal{P}(\overline{\mathbf{s}}_{t_0})$. In this way, one has existence and uniqueness on a closed time interval $\overline{\mathcal{T}} = [t_0, \overline{T}]$, for the given $\overline{\mathbf{s}}_{t_0} \in \mathcal{S}$ and for all $\mathbf{P}_{t_0} \in \overline{\mathcal{P}}(\overline{\mathbf{s}}_{t_0})$, along with the assurance that every $\mathbf{P}_{t_0} \in \overline{\mathcal{P}}(\overline{\mathbf{s}}_{t_0})$ is the covariance matrix of a random state $\mathbf{s}_{t_0}$ with the given mean $\overline{\mathbf{s}}_{t_0} \in \mathcal{S}$. Then one can conclude that a physically meaningful problem has been posed.

If one is given both $\overline{\mathbf{s}}_{t_0} \in \mathcal{S}$ and any particular $\mathbf{P}_{t_0} \in \mathcal{P}(\overline{\mathbf{s}}_{t_0})$, then existence and uniqueness are guaranteed on a closed time interval $\overline{\mathcal{T}} = [t_0, \overline{T}]$ by taking $\overline{\mathcal{P}} = \{\mathbf{P}_{t_0}\}$.

As a simple example of how the requirement that $\mathbf{P}_{t_0}$ is the covariance matrix of a random state with mean $\overline{\mathbf{s}}_{t_0}$ makes $\mathbf{P}_{t_0}$ depend on $\overline{\mathbf{s}}_{t_0}$, consider again the shallow-water system described at the end of Sec. 3. There it was shown that on the state space $\mathcal{S}_c$, solutions exist over arbitrarily long time intervals, for appropriate spatial discretizations. In terms of the variable u rather than α, the space $\mathcal{S}_c$ is described by the inequality

$$u_n^2 + 2c|u_n| + c^2 < \phi_n,$$

for $n = 1, \ldots, N$. Taking expectations gives

$$\mathcal{E}\left(u_n'\right)^2 + \overline{u}_n^2 + 2c\mathcal{E}|u_n| + c^2 < \overline{\phi}_n,$$

and since $|\overline{u}_n| < \mathcal{E}|u_n|$, this implies that

$$\mathcal{E}\left(u_n'\right)^2 < \overline{\phi}_n - \left(|\overline{u}_n| + c\right)^2. \tag{37}$$

Since the mean state is in $\mathcal{S}_c$, one has $\left(|\overline{u}_n| + c\right)^2 < \overline{\phi}_n$, so that the right-hand side of inequality (37) is indeed positive. This inequality is a restriction on the variance $\mathcal{E}\left(u_n'\right)^2$ in terms of $|\overline{u}_n|$ and $\overline{\phi}_n$ for every random state on $\mathcal{S}_c$, and the restriction becomes stronger as the mean state approaches the boundary of $\mathcal{S}_c$, that is, as the right-hand side of inequality (37) becomes small. For given $\overline{\phi}_n$, $n = 1, \ldots, N$, the restriction is mildest when the mean state is a state of rest, $\overline{u}_n = 0$ for $n = 1, \ldots, N$.

Returning to the general problem, suppose now that the set $\mathcal{P} = \mathcal{P}(\overline{\mathbf{s}}_{t_0})$ restricted by positive semidefiniteness and the geometry of $\mathcal{S}$ has been defined for each $\overline{\mathbf{s}}_{t_0} \in \mathcal{S}$. From inequality (36) and the fact that $\mathcal{S}$ is an open set it follows that $\mathcal{P}(\overline{\mathbf{s}}_{t_0})$ is an open set, for each $\overline{\mathbf{s}}_{t_0} \in \mathcal{S}$. A simple choice for the set $\overline{\mathcal{P}} = \overline{\mathcal{P}}(\overline{\mathbf{s}}_{t_0}) \subset \mathcal{P}(\overline{\mathbf{s}}_{t_0})$ of initial covariance matrices $\mathbf{P}_{t_0}$ is the set

$$\overline{\mathcal{P}} = \mathcal{P}_\mu = \{\mathbf{P} \in \mathcal{P} : \tfrac{1}{\mu}\mathbf{P} \in \mathcal{P}\}, \tag{38}$$

for any μ with $0 < \mu < 1$, which will be made for the sake of definiteness in Sec. 9, where a specific choice of $\overline{\mathcal{P}}$ is needed. Since $\mathcal{P}(\overline{\mathbf{s}}_{t_0})$ is an open set, $\mathcal{P}_\mu(\overline{\mathbf{s}}_{t_0})$ is also an open set. Note that both $\mathcal{P}(\overline{\mathbf{s}}_{t_0})$ and $\mathcal{P}_\mu(\overline{\mathbf{s}}_{t_0})$ contain the origin in $\mathbb{R}^{N\times N}$, for every $\overline{\mathbf{s}}_{t_0} \in \mathcal{S}$, since $\mathcal{S}$ is an open set.

In case the solution of the deterministic initial-value problem for Eq. (9) exists over arbitrarily long time intervals, one would like to know whether or not the solution of the initial-value problem for Eqs. (34) and (35) also exists over arbitrarily long time intervals, for each $\overline{\mathbf{s}}_{t_0} \in \mathcal{S}$ and $\mathbf{P}_{t_0} \in \mathcal{P}(\overline{\mathbf{s}}_{t_0})$. This important question is not addressed in the present article. Also, the question of how well the solution of Eqs. (34) and (35) approximates the mean and covariance matrix of the stochastic process defined in Sec. 4 is not considered here.

For given $\overline{\mathbf{s}}_{t_0} \in \mathcal{S}$ and all $\mathbf{P}_{t_0} \in \overline{\mathcal{P}}(\overline{\mathbf{s}}_{t_0})$, there exists on some closed time interval $\overline{\mathcal{T}} = [t_0, \overline{T}]$ a unique solution of Eqs. (34) and (35) satisfying the initial conditions. From this point onwards, $\overline{\mathbf{s}}_t$ and $\mathbf{P}_t$ (defined originally in Sec. 4) are redefined as this solution. The random vector $\mathbf{s}'_t = \mathbf{s}'_t(\omega)$ is also redefined, for each $\omega \in \Omega$ such that $\mathbf{s}_{t_0}(\omega) = \overline{\mathbf{s}}_{t_0} + \mathbf{s}'_{t_0}(\omega) \in \mathcal{S}$, as the unique solution on $\overline{\mathcal{T}}$ of the *(stochastic) perturbation equation*

$$\frac{d\mathbf{s}'}{dt} + \mathbf{F}(\overline{\mathbf{s}})\mathbf{s}' = \mathbf{0}, \tag{39}$$

corresponding to given initial condition $\mathbf{s}'_{t_0}(\omega) = \mathbf{s}_{t_0}(\omega) - \overline{\mathbf{s}}_{t_0}$. That there exists a unique solution on $\overline{\mathcal{T}}$ for each such ω is guaranteed by Assumption 2, since the perturbation equation is linear. Finally the stochastic process $\{\mathbf{s}_t : t \in \overline{\mathcal{T}}\}$ is defined, as the one whose sample paths are given by $\mathbf{s}_t(\omega) = \overline{\mathbf{s}}_t + \mathbf{s}'_t(\omega)$. The mean state of this second-order process is $\overline{\mathbf{s}}_t$, and $\mathbf{P}_t$ is its covariance matrix. Note that for this process,

$$\frac{d\mathbf{s}_t}{dt} = \frac{d\overline{\mathbf{s}}_t}{dt} + \frac{d\mathbf{s}'_t}{dt} \tag{40}$$

is continuous on $\overline{\mathcal{T}}$ wp1, since $d\overline{\mathbf{s}}_t/dt$ is continuous on $\overline{\mathcal{T}}$ and $d\mathbf{s}'_t/dt$ is continuous on $\overline{\mathcal{T}}$ wp1.

The stochastic perturbation equation has the form of the deterministic perturbation equation (11) for the original nonlinear dynamics. When $\mathbf{f}$ is nonlinear, the evolution of $\mathbf{s}'_t$ according to the stochastic perturbation equation depends through the Jacobian matrix on the mean state, and therefore also on the covariance matrix, since Eqs. (34) and (35) are fully coupled. Thus in the second-moment closure framework, to simulate the evolution of individual sample paths $\mathbf{s}_t(\omega)$ requires first solving for not only the mean state, but for the covariance matrix as well. This is one consequence of the presence of the nonlinear coupling term in the mean equation.

To close this section, it will be established that Eqs. (28) and (31) hold for the newly defined stochastic process. First, because the stochastic perturbation equation is linear, it has a fundamental matrix $\mathbf{M}_{t,t_0}$, not to be confused with that of the deterministic perturbation equation defined by Eq. (10). The fundamental matrix of the stochastic perturbation equation is defined for all $t \in \overline{\mathcal{T}}$ as the solution of the deterministic, linear equation

$$\frac{d\mathbf{M}_{t,t_0}}{dt} + \mathbf{F}(\overline{\mathbf{s}})\mathbf{M}_{t,t_0} = \mathbf{0} \tag{41}$$

corresponding to initial condition $\mathbf{M}_{t_0,t_0} = \mathbf{I}$, the $N \times N$ identity matrix. Therefore it expresses the solution of the stochastic perturbation equation directly in terms of random initial condition $\mathbf{s}'_{t_0}$:

$$\mathbf{s}'_t = \mathbf{M}_{t,t_0}\mathbf{s}'_{t_0}. \tag{42}$$

This fundamental matrix depends in general on the mean state, and therefore also on the covariance matrix, due to the nonlinear coupling term in the mean

equation. This dependence can be expressed fully as $\mathbf{M}_{t,\underline{t_0}} = \mathbf{M}_{t,t_0}(\overline{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0})$, since $\overline{\mathbf{s}}_{t_0}$ and $\mathbf{P}_{t_0}$ determine $\overline{\mathbf{s}}_t$ and $\mathbf{P}_t$ uniquely, for all $t \in \overline{\mathcal{T}}$. The fundamental matrix does not depend on the probability variable ω. Taking expectations in Eq. (42) therefore gives $\mathcal{E}\mathbf{s}'_t = \mathbf{0}$ since $\mathcal{E}\mathbf{s}'_{t_0} = \mathbf{0}$, and then taking expectations in Eq. (39) gives

$$\mathcal{E}\frac{d\mathbf{s}'}{dt} = \mathbf{0},$$

since the Jacobian matrix does not depend on ω. Taking expectations in Eq. (40) then yields Eq. (28).

It can be verified that the same fundamental matrix can be used to express the solution of the covariance evolution equation (35), as

$$\mathbf{P}_t = \mathbf{M}_{t,t_0}\mathbf{P}_{t_0}\mathbf{M}_{t,t_0}^T, \tag{43}$$

which is the *operator form* of the covariance evolution equation. Since $\mathbf{P}_{t_0} = \mathcal{E}\mathbf{s}'_{t_0}\mathbf{s}'^T_{t_0}$ is symmetric positive semidefinite, it follows from Eq. (43) that $\mathbf{P}_t$ is also symmetric positive semidefinite. From Eq. (42) one has

$$\mathbf{s}'_t\mathbf{s}'^T_t = \mathbf{M}_{t,t_0}\mathbf{s}'_{t_0}\mathbf{s}'^T_{t_0}\mathbf{M}_{t,t_0}^T,$$

and on taking expectations it follows that

$$\mathbf{P}_t = \mathcal{E}\mathbf{s}'_t\mathbf{s}'^T_t. \tag{44}$$

Postmultiplying Eq. (39) by $\mathbf{s}'^T$, then adding the transpose of the result to itself, then taking expectations and using Eq. (44), leads to

$$\mathcal{E}\frac{d\mathbf{s}'\mathbf{s}'^T}{dt} + \mathbf{F}(\overline{\mathbf{s}})\mathbf{P} + \mathbf{P}\mathbf{F}^T(\overline{\mathbf{s}}) = \mathbf{0}.$$

Comparing this result with Eq. (35) gives Eq. (31).

To summarize: Attention will now be focused on the mean equation (34), the covariance evolution equation (35), which can be written equivalently in operator form as Eq. (43), and the stochastic perturbation equation (39), which can be written equivalently in operator form as Eq. (42). The covariance matrix can be taken to be defined by Eq. (44). The fundamental matrix is defined by Eq. (41).

6 Energetic Consistency of the Second-Moment Closure Equations

The second-moment closure equations were derived without reference to Assumption 3 that total energy is conserved by the nonlinear dynamics (9). Assumption 3 will now be used to show that the second-moment closure equations are energetically consistent.

By Assumption 5, Eq. (17) can be differentiated twice with respect to each of the N state variables and then evaluated for any $\mathbf{s} \in \mathcal{S}$ and $t \in \mathcal{T}$. Doing so once gives

$$\mathbf{s}^T \frac{\partial \mathbf{f}}{\partial s_k} + \mathbf{e}^{kT} \mathbf{f} = 0, \tag{45}$$

for each $\mathbf{s} \in \mathcal{S}$ and $t \in \mathcal{T}$, and for $k = 1, \ldots, N$, where $\mathbf{e}^k$ denotes the k^{th} column of the $N \times N$ identity matrix. This can be written equivalently as

$$\mathbf{f}(\mathbf{s}) = -\mathbf{F}^T(\mathbf{s})\mathbf{s}, \tag{46}$$

which is a special relationship between $\mathbf{f}$ and its Jacobian matrix. Equation (46) implies in particular that if $\mathbf{0} \in \mathcal{S}$, then

$$\mathbf{f}(\mathbf{0}) = \mathbf{0}, \tag{47}$$

which means simply that the nonlinear dynamics are not externally forced, and that $\mathbf{s} = \mathbf{0}$ is a steady-state solution of the nonlinear dynamics. Recall, however, that typically $\mathbf{0} \notin \mathcal{S}$ for geophysical dynamics, since there are usually state variables that must be positive on physical grounds.

Differentiating Eq. (45) gives

$$\mathbf{s}^T \frac{\partial^2 \mathbf{f}}{\partial s_j \partial s_k} + \mathbf{e}^{jT} \frac{\partial \mathbf{f}}{\partial s_k} + \mathbf{e}^{kT} \frac{\partial \mathbf{f}}{\partial s_j} = 0, \tag{48}$$

for each $\mathbf{s} \in \mathcal{S}$ and $t \in \mathcal{T}$, and for $j, k = 1, \ldots, N$. The symmetric and anti-symmetric (skew-symmetric) parts of the Jacobian matrix, respectively $\mathbf{F}^s$ and $\mathbf{F}^a$, were defined following Eq. (13). Equation (48) can be rewritten in terms of $\mathbf{F}^s$, as

$$\mathbf{F}^s_{jk}(\mathbf{s}) = -\tfrac{1}{2}\mathbf{s}^T \frac{\partial^2 \mathbf{f}(\mathbf{s})}{\partial s_j \partial s_k}, \tag{49}$$

for each $\mathbf{s} \in \mathcal{S}$ and $t \in \mathcal{T}$, and for $j, k = 1, \ldots, N$. Energetic consistency of the second-moment closure equations is an immediate consequence of this special relationship between the symmetric part of the Jacobian matrix and the second partial derivatives of $\mathbf{f}$.

To see this, premultiply the mean equation (34) by $\overline{\mathbf{s}}^T$, and evaluate Eqs. (17) and (49) at $\mathbf{s} = \overline{\mathbf{s}} \in \mathcal{S}$, to find that

$$\tfrac{1}{2} \frac{d\overline{\mathbf{s}}^T \overline{\mathbf{s}}}{dt} - \sum_j \sum_k \mathbf{F}^s_{jk}(\overline{\mathbf{s}}) P_{jk} = 0. \tag{50}$$

Recall that the total variance V is the trace of the covariance matrix,

$$V = \operatorname{tr} \mathbf{P} = \mathcal{E}\mathbf{s}'^T \mathbf{s}',$$

where the second equality follows from Eq. (44). Applying the trace operator to the covariance evolution equation (35), and using the property that $\operatorname{tr} \mathbf{P}\mathbf{F}^T = \operatorname{tr} \mathbf{F}^T \mathbf{P}$, gives

$$\tfrac{1}{2} \frac{dV}{dt} + \operatorname{tr} \mathbf{F}^s(\overline{\mathbf{s}}) \mathbf{P} = 0. \tag{51}$$

This result follows also from the stochastic perturbation equation, as it must, by premultiplying Eq. (39) by $\mathbf{s}'^T$, then using the fact that $\mathbf{s}'^T(\omega)\mathbf{F}^a(\overline{\mathbf{s}})\mathbf{s}'(\omega) = 0$ for each $\omega \in \Omega$, since $\mathbf{F}^a$ is skew-symmetric and $\mathbf{s}'$ is real, and then taking expectations. Adding Eqs. (50) and (51) gives

$$\frac{d(\overline{\mathbf{s}}^T\overline{\mathbf{s}} + V)}{dt} = 0, \tag{52}$$

which is the statement (20) of energetic consistency for the second-moment closure equations. It implies in particular the bounds (21) and (22) on their solutions.

7 The Role of the Symmetric Part of the Jacobian Matrix

Consider for a moment the case of linear, conservative dynamics. If $\mathbf{f}(\mathbf{s}, t)$ is linear in $\mathbf{s}$, then the Jacobian matrix $\mathbf{F} = \mathbf{F}(\mathbf{s}, t)$ is independent of $\mathbf{s}$, and the second partial derivatives of $\mathbf{f}$ with respect to the state variables all vanish. Thus from Eq. (49) it follows that $\mathbf{F}^s = \mathbf{0}$, and therefore from Eq. (46) one has simply $\mathbf{f}(\mathbf{s}) = \mathbf{F}^a\mathbf{s}$, with $\mathbf{F}^a$ independent of $\mathbf{s}$. The mean and covariance evolution equations, already simple for linear dynamics in general, simplify still further for conservative dynamics.

That the Jacobian matrix has a symmetric part $\mathbf{F}^s$ is one consequence of nonlinearity. Moreover, Eqs. (50) and (51) show that the exchange of energy between the mean state and the stochastic perturbations, which leads to the exact balance (52), occurs solely through the symmetric part of the Jacobian matrix. Equation (51) shows also that $\mathbf{F}^s$ directly controls the growth and/or decay of the uncertainty measured by the total variance V.

This section gives an overview of the role that $\mathbf{F}^s$ plays in controlling the behavior of the total variance, and therefore in determining how the amount of energy exchanged with the mean state changes through time. Section 8 will then examine in more depth some of the special properties that $\mathbf{F}^s$ has in this role, which are properties that result from conservation of total energy.

According to Eq. (43), the solution $\mathbf{P}$ of the covariance evolution equation is symmetric positive semidefinite, with rank not exceeding that of $\mathbf{P}_{t_0}$. Therefore $\mathbf{P}$ has eigendecomposition

$$\mathbf{P} = \mathbf{W}\mathbf{\Sigma}^2\mathbf{W}^T,$$

where $\mathbf{\Sigma}^2$ is the diagonal matrix of non-negative eigenvalues $\sigma_1^2, \ldots, \sigma_L^2$, where $\operatorname{rank}\mathbf{P} \le L \le N$, and where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_L]$ is the $N \times L$ matrix of normalized real eigenvectors $\mathbf{w}_l$, $\mathbf{w}_l^T\mathbf{w}_l = 1$ for $l = 1, \ldots, L$. The eigenvalues and eigenvectors depend on time. It follows from the eigendecomposition that

$$V = \operatorname{tr}\mathbf{P} = \sum_{l=1}^{L} \sigma_l^2, \tag{53}$$

and that

$$\operatorname{tr} \mathbf{F}^s(\overline{\mathbf{s}})\mathbf{P} = \sum_{l=1}^{L} \sigma_l^2 \mathbf{w}_l^T \mathbf{F}^s(\overline{\mathbf{s}})\mathbf{w}_l. \tag{54}$$

Substituting into Eq. (51) gives

$$\sum_{l=1}^{L} \left[\tfrac{1}{2}\frac{d}{dt} + \mathbf{w}_l^T \mathbf{F}^s(\overline{\mathbf{s}})\mathbf{w}_l \right] \sigma_l^2 = 0. \tag{55}$$

Equation (55) says that the presence in $\mathbf{P}$ of an eigenvector $\mathbf{w}_l$ with nonzero eigenvalue acts to increase (decrease) uncertainty if $\mathbf{w}_l^T \mathbf{F}^s(\overline{\mathbf{s}})\mathbf{w}_l < 0$ ($\mathbf{w}_l^T \mathbf{F}^s(\overline{\mathbf{s}})\mathbf{w}_l > 0$). In particular, if $\mathbf{F}^s(\mathbf{s}, t)$ happens to be negative semidefinite, for all $\mathbf{s} \in \mathcal{S}$ and $\mathbf{t} \in \mathcal{T}$, then the total variance V is monotone nondecreasing, $dV/dt \geq 0$, and hence the total energy of the mean state is monotone nonincreasing, $d\overline{\mathbf{s}}^T\overline{\mathbf{s}}/dt \leq 0$, independently of the initial condition $(\overline{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0})$, for as long as the solution $(\overline{\mathbf{s}}, \mathbf{P})$ exists. Thus dynamics with negative semidefinite $\mathbf{F}^s$ constitute a worst case for predictability: the uncertainty V can only increase with time, or at best hold constant at times. That this should be a worst case is to be expected already from the deterministic initial-value problem: the deterministic perturbation equation (11) gives

$$\tfrac{1}{2}\frac{d\mathbf{q}^T\mathbf{q}}{dt} + \mathbf{q}^T \mathbf{F}^s(\mathbf{s}_t, t)\mathbf{q} = 0$$

for perturbations of the deterministic trajectory $\mathbf{s}_t$, and so the size $\mathbf{q}^T\mathbf{q}$ of the perturbation is monotone nondecreasing, regardless of the initial perturbation, if $\mathbf{F}^s$ is negative semidefinite on $\mathcal{S} \times \mathcal{T}$. Similarly, if $\mathbf{F}^s$ is positive semidefinite on $\mathcal{S} \times \mathcal{T}$, then $dV/dt \leq 0$ and $d\overline{\mathbf{s}}^T\overline{\mathbf{s}}/dt \geq 0$, independently of initial condition $(\overline{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0})$, for as long as the solution $(\overline{\mathbf{s}}, \mathbf{P})$ exists. The dynamics in this case would be eminently predictable.

By Assumptions 2 and 4, $\mathbf{F}^s$ is bounded on $\mathcal{S} \times \mathcal{T}$, and therefore the eigenvalues of $\mathbf{F}^s$ are bounded on $\mathcal{S} \times \mathcal{T}$. The eigenvalues are real, since $\mathbf{F}^s$ is symmetric. Denote the largest and smallest eigenvalues of $\mathbf{F}^s$ by $\lambda_{\max}(\mathbf{F}^s)$ and $\lambda_{\min}(\mathbf{F}^s)$, respectively. In the next section it will be shown that, for conservative dynamics,

$$\lambda_{\min}(\mathbf{F}^s(\mathbf{s}, t)) \leq 0 \leq \lambda_{\max}(\mathbf{F}^s(\mathbf{s}, t)), \tag{56}$$

at each point $\mathbf{s} \in \mathcal{S}$ and $t \in \mathcal{T}$. This means that $\mathbf{F}^s$ cannot be either positive or negative definite, anywhere on $\mathcal{S} \times \mathcal{T}$, although it does not rule out positive or negative semidefiniteness. Also it will be shown that at each point $\mathbf{s}$ in any open set $\mathcal{S}_*$ on which $\mathbf{f}$ is *genuinely nonlinear* at time t, as defined there, the inequalities in (56) are strict at time t:

$$\lambda_{\min}(\mathbf{F}^s(\mathbf{s}, t)) < 0 < \lambda_{\max}(\mathbf{F}^s(\mathbf{s}, t)), \tag{57}$$

which means that $\mathbf{F}^s(\mathbf{s}, t)$ has at least one positive and one negative eigenvalue at time t, at every point $\mathbf{s} \in \mathcal{S}_*$. Thus if $\mathbf{F}^s$ is genuinely nonlinear on all of $\mathcal{S}$, for

all $t \in \mathcal{T}$, then $\mathbf{F}^s$ cannot be either positive or negative semidefinite, anywhere on $\mathcal{S} \times \mathcal{T}$: there is potential for both growth and decay of uncertainty, at all times, regardless of where the mean state happens to be in $\mathcal{S}$. Equation (55) shows in the genuinely nonlinear case that whether growth, decay, or neither actually occurs at a particular time depends on the eigenstructure of $\mathbf{P}$, relative to that of $\mathbf{F}^s(\overline{\mathbf{s}})$, at that time.

Now denote by $\mathbf{u}_{\max} = \mathbf{u}_{\max}(\mathbf{F}^s(\overline{\mathbf{s}}, t))$ and $\mathbf{u}_{\min} = \mathbf{u}_{\min}(\mathbf{F}^s(\overline{\mathbf{s}}, t))$, respectively, the eigenvectors of $\mathbf{F}^s$ corresponding to eigenvalues $\lambda_{\max} = \lambda_{\max}(\mathbf{F}^s(\overline{\mathbf{s}}, t))$ and $\lambda_{\min} = \lambda_{\min}(\mathbf{F}^s(\overline{\mathbf{s}}, t))$. It follows from Eqs. (53) and (54) that

$$\min_{\mathbf{P}} \frac{\operatorname{tr} \mathbf{F}^s(\overline{\mathbf{s}}, t)\mathbf{P}}{\operatorname{tr} \mathbf{P}} = \lambda_{\min},$$

where the minimization is over all symmetric positive semidefinite matrices $\mathbf{P}$, and that the minimum is attained for $\mathbf{P} = \mathbf{P}_{\min} = \gamma \mathbf{u}_{\min} \mathbf{u}_{\min}^T$, with γ an arbitrary positive constant. Furthermore, $\mathbf{P}_{\min} \in \mathcal{P}(\overline{\mathbf{s}})$ for all γ small enough, where $\mathcal{P}$ is the set defined in Sec. 5, since $\mathcal{P}$ is an open set containing the origin in $\mathbb{R}^{N \times N}$. Also, $\mathbf{P}_{\min} \in \mathcal{P}_\mu(\overline{\mathbf{s}})$ for each $\mu \in (0, 1)$, by taking γ still smaller, where $\mathcal{P}_\mu$ is the set defined in Eq. (38). Thus the minimum is achieved for matrices in $\mathcal{P}(\overline{\mathbf{s}})$ and in $\mathcal{P}_\mu(\overline{\mathbf{s}})$. Similarly,

$$\max_{\mathbf{P}} \frac{\operatorname{tr} \mathbf{F}^s(\overline{\mathbf{s}}, t)\mathbf{P}}{\operatorname{tr} \mathbf{P}} = \lambda_{\max},$$

and the maximum is attained for $\mathbf{P} = \mathbf{P}_{\max} = \delta \mathbf{u}_{\max} \mathbf{u}_{\max}^T$, with δ an arbitrary positive constant. Rewriting Eq. (51) as

$$\tfrac{1}{2} \frac{1}{V} \frac{dV}{dt} + \frac{\operatorname{tr} \mathbf{F}^s(\overline{\mathbf{s}}, t)\mathbf{P}}{\operatorname{tr} \mathbf{P}} = 0,$$

then shows that $-2\lambda_{\min}$ ($+2\lambda_{\max}$) is the maximum instantaneous relative rate of increase (decrease) of uncertainty, and is attained for $\mathbf{P} = \mathbf{P}_{\min}$ ($\mathbf{P} = \mathbf{P}_{\max}$).

The presence of both positive and negative eigenvalues of $\mathbf{F}^s$ can lead to complex behavior for the total variance V as a function of time, and therefore also for the behavior of the total energy of the mean state as a function of time, which according to Eq. (52) mirrors precisely that of $V/2$. In general both growth and decay of total variance can occur, and therefore dV/dt can also vanish instantaneously. For instance, one can check that $dV/dt = 0$ at

$$\mathbf{P} = \lambda_{\max} \mathbf{u}_{\min} \mathbf{u}_{\min}^T - \lambda_{\min} \mathbf{u}_{\max} \mathbf{u}_{\max}^T.$$

8 Genuine Nonlinearity and Essential Linearity

Inequality (56) is readily established. First observe that Eqs. (17) and (46) together imply that

$$\mathbf{s}^T \mathbf{F}^s(\mathbf{s}) \mathbf{s} = 0, \tag{58}$$

for every $\mathbf{s} \in \mathcal{S}$ and $t \in \mathcal{T}$, since $\mathbf{F}^a$ is skew-symmetric. Also note for later reference that if $\mathbf{0} \in \mathcal{S}$, then Eq. (49) gives

$$\mathbf{F}^s(\mathbf{0}) = \mathbf{0}, \tag{59}$$

although again one should recall that typically $\mathbf{0} \notin \mathcal{S}$ for models of ocean and atmospheric dynamics.

Equation (58) implies that at each point $\mathbf{s} \in \mathcal{S}$, one has the following alternative, at each fixed time $t \in \mathcal{T}$: either $\mathbf{F}^s(\mathbf{s})$ has at least one positive and one negative eigenvalue, or else $\mathbf{s}$ is a null-vector of $\mathbf{F}^s(\mathbf{s})$,

$$\mathbf{F}^s(\mathbf{s})\mathbf{s} = \mathbf{0}. \tag{60}$$

To see this, let

$$\mathbf{F}^s(\mathbf{s}) = \mathbf{U}(\mathbf{s})\mathbf{\Lambda}(\mathbf{s})\mathbf{U}^T(\mathbf{s})$$

be the eigendecomposition of $\mathbf{F}^s$, where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_N$, all of which are real since $\mathbf{F}^s$ is real and symmetric, and where $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_N]$ is the matrix of normalized real eigenvectors $\mathbf{u}_l$, $\mathbf{u}_l^T\mathbf{u}_l = 1$ for $l = 1, \ldots, N$. Then Eq. (58) can be rewritten as

$$\sum_{l=1}^{N} \lambda_l(\mathbf{s})[\mathbf{s}^T\mathbf{u}_l(\mathbf{s})]^2 = 0.$$

Either all the terms in this sum vanish or there is at least one positive and one negative term, equivalently, at least one positive and one negative eigenvalue. The condition that all the terms vanish can be expressed as $\mathbf{\Lambda}(\mathbf{s})\mathbf{U}^T(\mathbf{s})\mathbf{s} = \mathbf{0}$, which is equivalent to Eq. (60) since $\mathbf{U}(\mathbf{s})$ is nonsingular. Thus the statement of alternatives has been demonstrated.

Inequality (56) holds at each point $\mathbf{s} \in \mathcal{S}$, $t \in \mathcal{T}$ where $\mathbf{F}^s(\mathbf{s}, t)$ has a null-vector. Therefore it holds at each $\mathbf{s} \in \mathcal{S}$, $t \in \mathcal{T}$ where the second alternative condition, Eq. (60), holds. Furthermore, inequality (57) holds at each point $\mathbf{s} \in \mathcal{S}$, $t \in \mathcal{T}$ where the first alternative condition, that $\mathbf{F}^s(\mathbf{s}, t)$ has at least one positive and one negative eigenvalue, holds. Therefore inequality (56) holds for all $\mathbf{s} \in \mathcal{S}$ and $t \in \mathcal{T}$. The rest of this section will distinguish the two alternatives more tangibly, and examples will be given at the end of the section.

The condition expressed by Eq. (60) can be expressed equivalently in a way that clarifies when it occurs and also allows one to check for its occurrence essentially by inspection of $\mathbf{f}$, without even calculating $\mathbf{F}^s$. This is done by first introducing the polar coordinate $\rho = (\mathbf{s}^T\mathbf{s})^{1/2}$, so that for each $\mathbf{s} \in \mathcal{S}$ one has $\mathbf{s} = \rho\mathbf{c}$ with $\mathbf{c}$ on the unit hypersphere $\mathbf{c}^T\mathbf{c} = 1$ in $\mathbb{R}^N$. Any scalar, vector, or matrix function $\phi = \phi(\mathbf{s}) = \phi(\mathbf{s}, t)$ that is continuously differentiable with respect to $\mathbf{s} \in \mathcal{S}$ for all $t \in \mathcal{T}$ also has a continuous derivative with respect to ρ, for all $\mathbf{s} \in \mathcal{S}$ and $t \in \mathcal{T}$, and

$$\frac{\partial \phi}{\partial \rho} = \sum_{l=1}^{N} \frac{\partial \phi}{\partial s_l}\frac{\partial s_l}{\partial \rho} = \sum_{l=1}^{N} \frac{\partial \phi}{\partial s_l} c_l,$$

so that

$$\rho\frac{\partial\phi}{\partial\rho}=\sum_{l=1}^{N}\frac{\partial\phi}{\partial s_l}s_l.$$

By Assumption 2, ϕ can be taken to be $\mathbf{f}$, in which case this relationship reads

$$\rho\frac{\partial\mathbf{f}(\mathbf{s})}{\partial\rho}=\mathbf{F}(\mathbf{s})\mathbf{s},$$

since $\partial\mathbf{f}(\mathbf{s})/\partial s_l$ is by definition the l^{th} column of the Jacobian matrix $\mathbf{F}(\mathbf{s})$. Using Eq. (46), one then has

$$\mathbf{F}^s(\mathbf{s})\mathbf{s}=\tfrac{1}{2}\left[\mathbf{F}(\mathbf{s})+\mathbf{F}^T(\mathbf{s})\right]\mathbf{s}=\tfrac{1}{2}\left[\rho\frac{\partial\mathbf{f}(\mathbf{s})}{\partial\rho}-\mathbf{f}(\mathbf{s})\right]=\tfrac{1}{2}\rho^2\frac{\partial\rho^{-1}\mathbf{f}(\mathbf{s})}{\partial\rho},\qquad(61)$$

for every $\mathbf{s}\in\mathcal{S}$ and $t\in\mathcal{T}$.

Thus the second alternative condition, Eq. (60), is equivalent to

$$\mathbf{f}(\mathbf{s})=\rho\frac{\partial\mathbf{f}(\mathbf{s})}{\partial\rho}.\qquad(62)$$

It is always the second alternative that holds at the origin, if $\mathbf{0}\in\mathcal{S}$, as seen either by setting $\mathbf{s}=\mathbf{0}$ in Eq. (60) or by setting $\rho=0$ in Eq. (62) and using Eq. (47). Away from the origin, Eq. (61) says that the second alternative condition is equivalent to

$$\frac{\partial\rho^{-1}\mathbf{f}(\mathbf{s})}{\partial\rho}=\mathbf{0}.\qquad(63)$$

Substituting Eq. (62) into Eq. (63) shows that, away from the origin, the second alternative condition is equivalent simply to

$$\frac{\partial^2\mathbf{f}(\mathbf{s})}{\partial\rho^2}=\mathbf{0}.\qquad(64)$$

If Eq. (64) holds not just at a point $\mathbf{s}\in\mathcal{S}$, but for all points in an open set $\mathcal{S}_*\subseteq\mathcal{S}$, then $\mathbf{f}$ is linear as a function of ρ for all $\mathbf{s}\in\mathcal{S}_*$. By Assumption 5, $\partial^2\mathbf{f}/\partial\rho^2$ is continuous on the state space $\mathcal{S}$ for each $t\in\mathcal{T}$, and therefore if Eq. (64) holds on an open set $\mathcal{S}_*\subseteq\mathcal{S}$, then it holds also on $\overline{\mathcal{S}_*}\cap\mathcal{S}$, where $\overline{\mathcal{S}_*}$ denotes the closure of $\mathcal{S}_*$ in $\mathbb{R}^N$. If $\mathbf{f}$ is linear in all the state variables, on an open set $\mathcal{S}_*\subseteq\mathcal{S}$, then certainly $\mathbf{f}$ is linear in ρ throughout $\mathcal{S}_*$, and not only does the second alternative therefore hold on $\mathcal{S}_*$, but it was seen at the beginning of Sec. 7 that then in fact $\mathbf{F}^s=\mathbf{0}$ on $\mathcal{S}_*$. It is possible for $\mathbf{f}$ to be linear in ρ on an open set $\mathcal{S}_*\subseteq\mathcal{S}$ without being linear in any of the state variables there. This is the case when $\mathbf{f}(\mathbf{s})=\rho\mathbf{g}(\mathbf{c})$ for $\mathbf{s}\in\mathcal{S}_*$, for some function $\mathbf{g}$ whose first partial derivatives with respect to all the $c_l, l=1,\ldots,N$, vanish nowhere for $\mathbf{s}=\rho\mathbf{c}\in\mathcal{S}_*$.

In case condition (64) holds at every point $\mathbf{s}$ in an open set $\mathcal{S}_*=\mathcal{S}_*(t)\subseteq\mathcal{S}$ at a particular time $t\in\mathcal{T}$, then $\mathbf{f}=\mathbf{f}(\mathbf{s},t)$ will be said to be *essentially linear* on

$\mathcal{S}_*(t)$. The equivalence between conditions (60) and (64), together with Eq. (46), implies that $\mathbf{f}(\mathbf{s},t)$ is essentially linear on an open set $\mathcal{S}_*(t)$ if, and only if,

$$\mathbf{f}(\mathbf{s},t) = \mathbf{F}^a(\mathbf{s},t)\mathbf{s}, \tag{65}$$

for all $\mathbf{s} \in \mathcal{S}_*(t)$. The discussion at the beginning of Sec. 7 shows that if $\mathbf{f}(\mathbf{s},t)$ is linear in all the state variables on an open set $\mathcal{S}_*(t) \subseteq \mathcal{S}$, then Eq. (65) holds with $\mathbf{F}^a(\mathbf{s},t)$ independent of $\mathbf{s}$ on $\mathcal{S}_*(t)$. Thus essential linearity, as defined here, amounts to generalizing the notion of linearity in such a way that one still has $\mathbf{f} = \mathbf{F}^a\mathbf{s}$, but with $\mathbf{F}^a$ depending on $\mathbf{s}$.

In case condition (64) holds at *no* point $\mathbf{s}$ in an open set $\mathcal{S}_* = \mathcal{S}_*(t) \subseteq \mathcal{S}$ at a particular time $t \in \mathcal{T}$, then $\mathbf{f} = \mathbf{f}(\mathbf{s},t)$ will be said to be *genuinely nonlinear* on $\mathcal{S}_*(t)$. With this definition, the original statement of alternatives can finally be rephrased, for open sets instead of points, as follows: $\mathbf{F}^s(\mathbf{s},t)$ has at least one positive and one negative eigenvalue at each point $\mathbf{s}$ in an open set $\mathcal{S}_*(t) \subseteq \mathcal{S}$ if, and only if, $\mathbf{f}(\mathbf{s},t)$ is genuinely nonlinear on $\mathcal{S}_*(t)$. It has also been shown that an open set $\mathcal{S}_*(t) \subseteq \mathcal{S}$ on which $\mathbf{f}(\mathbf{s},t)$ is genuinely nonlinear cannot contain the origin $\mathbf{s} = \mathbf{0}$, since Eq. (60) holds at the origin. If $\mathbf{0} \notin \mathcal{S}$, as is typical for geophysical problems, then it is possible for $\mathbf{f}$ to be genuinely nonlinear on all of $\mathcal{S}$, for all time $t \in \mathcal{T}$. It is not difficult to show that for reasonable discretizations of the one-dimensional shallow-water system considered in Secs. 3, 4 and 5, $\mathbf{f}$ is genuinely nonlinear on all of the state space $\mathcal{S}_c$, for arbitrarily long time intervals $\mathcal{T}$.

To illustrate the ideas of genuine nonlinearity and essential linearity in the simplest setting, consider the case $N = 2$. It follows from Eq. (17) that $\mathbf{f}$ has the form

$$\mathbf{f}(\mathbf{s},t) = \beta(\mathbf{s},t) \begin{bmatrix} s_2 \\ -s_1 \end{bmatrix} = \rho\beta(\mathbf{s},t) \begin{bmatrix} c_2 \\ -c_1 \end{bmatrix},$$

for some scalar function β, thus generalizing the first example of Sec. 3. It is immediate that, away from the origin, $\mathbf{f}$ is essentially linear precisely on those open sets $\mathcal{S}_*(t)$ on which β is independent of ρ at time t, that is, on which β is a function only of the ratio s_1/s_2 at time t. That Eq. (65) does indeed hold on every such set can be verified directly, by calculating $\mathbf{F}^a$. One finds that

$$\mathbf{F}^a = \left(\beta + \tfrac{1}{2}\rho\frac{\partial\beta}{\partial\rho}\right) \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix},$$

so that

$$\mathbf{F}^a\mathbf{s} = \left(\beta + \tfrac{1}{2}\rho\frac{\partial\beta}{\partial\rho}\right) \begin{bmatrix} s_2 \\ -s_1 \end{bmatrix},$$

and so Eq. (65) does hold where $\partial\beta/\partial\rho = 0$, as well as at the origin. Similarly, it is immediate that, away from the origin, $\mathbf{f}$ is genuinely nonlinear precisely on those open sets $\mathcal{S}_*(t)$ on which $\partial\beta/\partial\rho$ vanishes nowhere at time t. The statement of alternatives says that on such a set, $\mathbf{F}^s$ must have at least one positive and one negative eigenvalue, which for $N = 2$ means that $\mathbf{F}^s$ must

have exactly one positive and one negative eigenvalue, and therefore must have a negative determinant, $\det \mathbf{F}^s < 0$. One can calculate directly that

$$\det \mathbf{F}^s = -\tfrac{1}{4}\rho^2 \left(\frac{\partial \beta}{\partial \rho}\right)^2,$$

which verifies the statement.

Finally, although it is not usually the case that $\mathbf{0} \in \mathcal{S}$ for geophysical problems, it is instructive to consider dynamical behavior near the origin in case $\mathbf{0} \in \mathcal{S}$, particularly in light of the ideas of genuine nonlinearity and essential linearity. Recall from Eq. (47) that $\mathbf{s} = \mathbf{0}$ is a steady-state solution of the original, deterministic nonlinear dynamics (9).

Consider first the case in which $\mathbf{f}$ is essentially linear on an open set $\mathcal{S}_*(t) \subseteq \mathcal{S}$ with $\mathbf{0} \in \mathcal{S}_*(t)$. Thus $\mathbf{f}$ is linear in ρ on $\mathcal{S}_*(t)$, and hence $\mathbf{f}$ is linear in ρ near the origin along each coordinate axis. Therefore

$$f_k(\epsilon \mathbf{e}^l, t) = \alpha_k^l(t)|\epsilon|, \tag{66}$$

for some scalars α_k^l, $k, l = 1, \ldots, N$, and for all ϵ small enough, where $\mathbf{e}^l$ denotes the l^{th} column of the $N \times N$ identity matrix. This follows from the fact that $\rho = (\mathbf{s}^T\mathbf{s})^{1/2} = |\epsilon|$ for $\mathbf{s} = \epsilon \mathbf{e}^l$. By Assumption 2, $\partial f_k/\partial s_l$ exists and is continuous at the origin, and since $|\epsilon|$ is not differentiable at $\epsilon = 0$, it follows from Eq. (66) that $\alpha_k^l(t) = 0$ for $k, l = 1, \ldots, N$. Therefore $\mathbf{f}(\mathbf{s}, t) = \mathbf{0}$ for all $\mathbf{s} \in \mathcal{S}_*(t)$: essentially linear dynamics near the origin are trivial dynamics. Further, if $\mathbf{f}$ is essentially linear on all of $\mathcal{S}$, for all $t \in \mathcal{T}$, and if $\mathbf{0} \in \mathcal{S}$, what has just been shown is that then $\mathbf{f} = \mathbf{0}$ on $\mathcal{S} \times \mathcal{T}$.

Now consider dynamics near the origin for general $\mathbf{f}$. If $\mathbf{0} \in \mathcal{S}$, one has that $\epsilon \mathbf{e}^l \in \mathcal{S}$, for $l = 1, \ldots, N$ and for all ϵ small enough. Evaluating Eq. (49) at $\mathbf{s} = \epsilon \mathbf{e}^l$ and using Eq. (59) gives

$$\frac{1}{\epsilon}\left[\mathbf{F}_{jk}^s(\epsilon \mathbf{e}^l, t) - \mathbf{F}_{jk}^s(\mathbf{0}, t)\right] = -\tfrac{1}{2}\mathbf{e}^{lT}\frac{\partial^2 \mathbf{f}(\epsilon \mathbf{e}^l, t)}{\partial s_j \partial s_k},$$

and taking the limit here as $\epsilon \to 0$ then gives

$$\frac{\partial \mathbf{F}_{jk}^s(\mathbf{0}, t)}{\partial s_l} = -\tfrac{1}{2}\frac{\partial^2 f_l(\mathbf{0}, t)}{\partial s_j \partial s_k}.$$

In view of Eq. (47), at $\overline{\mathbf{s}} = \mathbf{0}$ the mean equation (34) therefore reads

$$\frac{d\overline{s}_l}{dt} - \sum_j \sum_k \frac{\partial \mathbf{F}_{jk}^s(\mathbf{0}, t)}{\partial s_l} P_{jk} = \mathbf{0}, \tag{67}$$

for $l = 1, \ldots, N$. Thus the behavior of the mean state $\overline{\mathbf{s}}$ at $\overline{\mathbf{s}} = \mathbf{0}$ depends on the symmetric matrices $\partial \mathbf{F}^s(\mathbf{0}, t)/\partial s_l$, $l = 1, \ldots, N$. In particular, even though $\mathbf{s} = \mathbf{0}$ is a steady-state solution of the deterministic nonlinear dynamics, $\overline{\mathbf{s}} = \mathbf{0}$ is not necessarily a steady-state solution of the mean equation. The reason for this

is the presence of uncertainty in the stochastic dynamics, which is manifested through the covariance matrix in the nonlinear coupling term in Eq. (67).

In view of Eq. (59), at $\overline{\mathbf{s}} = \mathbf{0}$ the covariance evolution equation (35) simplifies to

$$\frac{d\mathbf{P}}{dt} + \mathbf{F}^a(\mathbf{0}, t)\mathbf{P} - \mathbf{P}\mathbf{F}^a(\mathbf{0}, t) = \mathbf{0}. \tag{68}$$

Thus at $\overline{\mathbf{s}} = \mathbf{0}$ the covariance matrix evolves in an energetically neutral way, $dV/dt = 0$, and the evolution of the mean state at $\overline{\mathbf{s}} = \mathbf{0}$ still depends on this covariance evolution, through the nonlinear coupling term in Eq. (67). In particular, it is possible for the mean state to remain zero for a period of time, not just instantaneously, and subsequently to regain energy. That is to say, although it is possible for all of the energy of the mean state to have been extracted at some particular time $t = \tau$, and therefore for the upper bound in Eq. (21) actually to be attained at time τ, it is also possible for the mean state to regain energy after time τ, through interaction with the covariance matrix in the nonlinear coupling term. This effect is not present if the nonlinear coupling term is neglected.

9 Bounds on the Growth of Relative Uncertainty

It was shown in Sec. 7 that the instantaneous relative rate of change of the total variance V_t satisfies the bounds

$$-2\lambda_{\max}(\mathbf{F}^s(\overline{\mathbf{s}}_t, t)) \leq \frac{1}{V_t}\frac{dV_t}{dt} \leq -2\lambda_{\min}(\mathbf{F}^s(\overline{\mathbf{s}}_t, t)),$$

with each bound attainable by rank-one matrices $\mathbf{P}_t \in \mathcal{P}_\mu(\overline{\mathbf{s}}_t) \subset \mathcal{P}(\overline{\mathbf{s}}_t)$, for any given $\mu \in (0, 1)$, where $\mathcal{P}$ and $\mathcal{P}_\mu$ are the sets defined in Sec. 5. It was shown in Sec. 8 that $\lambda_{\min} \leq 0$ and $\lambda_{\max} \geq 0$, and furthermore that if $\overline{\mathbf{s}}_t$ is in an open set on which $\mathbf{f}(\overline{\mathbf{s}}_t, t)$ is genuinely nonlinear, then $\lambda_{\min} < 0$ and $\lambda_{\max} > 0$. Both $\lambda_{\max}$ and $|\lambda_{\min}|$ can be large, although finite, since no assumption has been introduced that would otherwise limit these values. Thus not only is it possible to have both growth and decay of total variance, at any instant of time it is also possible for the relative rate of growth or decay to be large.

On the other hand, inequalities (21) and (22) show that V_t cannot grow unboundedly. That is to say, although the total variance can grow rapidly at particular *instants* of time, growth over any *interval* of time is strictly limited, due to the energetic consistency of the mean and covariance evolution equations which was demonstrated in Sec. 6. The present section examines the growth of V_t, relative to V_{t_0}, over every interval of time for which the solution of the second-moment closure equations exists, thus providing time-independent bounds on the relative uncertainty V_t/V_{t_0}. The latter part of the discussion concerns inequalities (21) and (22) only, for the given state space $\mathcal{S}$, without specific reference to the particular problem whose solutions satisfy the inequalities. Therefore the time-independent bounds obtained for V_t/V_{t_0} apply as well to the original stochastic process defined in Sec. 4, for as long as it exists.

Consider first the behavior of solutions of the second-moment closure equations when the nonlinear coupling term in the mean equation is neglected, so that the fundamental matrix $\mathbf{M}_{t,t_0} = \mathbf{M}_{t,t_0}(\overline{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0})$ introduced in Sec. 5 becomes a function of $\overline{\mathbf{s}}_{t_0}$ alone, $\mathbf{M}_{t,t_0} = \mathbf{M}_{t,t_0}(\overline{\mathbf{s}}_{t_0})$. Denote by $\sigma_{t,t_0} = \sigma_{t,t_0}(\overline{\mathbf{s}}_{t_0})$ the largest singular value of $\mathbf{M}_{t,t_0}$, and denote by $\mathbf{v}_{t_0} = \mathbf{v}_{t_0}(\overline{\mathbf{s}}_{t_0})$ the corresponding right singular vector, normalized so that $\mathbf{v}_{t_0}^T \mathbf{v}_{t_0} = 1$. The corresponding normalized left singular vector, $\mathbf{u}_t = \mathbf{u}_t(\overline{\mathbf{s}}_{t_0})$, is then given by

$$\mathbf{u}_t = \frac{1}{\sigma_{t,t_0}} \mathbf{M}_{t,t_0} \mathbf{v}_{t_0}.$$

Now take

$$\mathbf{P}_{t_0} = V_{t_0} \mathbf{v}_{t_0} \mathbf{v}_{t_0}^T, \tag{69}$$

so that $V_{t_0} = \operatorname{tr} \mathbf{P}_{t_0}$, and use Eq. (43) to obtain

$$\frac{V_t}{V_{t_0}} = \frac{\operatorname{tr} \mathbf{M}_{t,t_0} \mathbf{P}_{t_0} \mathbf{M}_{t,t_0}^T}{\operatorname{tr} \mathbf{P}_{t_0}} = \frac{\operatorname{tr} \sigma_{t,t_0}^2 V_{t_0} \mathbf{u}_t \mathbf{u}_t^T}{\operatorname{tr} \mathbf{P}_{t_0}} = \sigma_{t,t_0}^2.$$

The largest singular value σ_{t,t_0} of any matrix $\mathbf{M}_{t,t_0}$ also has the property that

$$\frac{\operatorname{tr} \mathbf{M}_{t,t_0} \mathbf{P} \mathbf{M}_{t,t_0}^T}{\operatorname{tr} \mathbf{P}} \leq \sigma_{t,t_0}^2, \tag{70}$$

for every symmetric positive semidefinite matrix $\mathbf{P}$. This leads to the usual result that

$$\frac{V_t}{V_{t_0}} \leq \sigma_{t,t_0}^2(\overline{\mathbf{s}}_{t_0}), \tag{71}$$

with equality holding for the choice of $\mathbf{P}_{t_0}$ given in Eq. (69). For this choice it is guaranteed that $\mathbf{P}_{t_0} \in \mathcal{P}_\mu(\overline{\mathbf{s}}_{t_0}) \subset \mathcal{P}(\overline{\mathbf{s}}_{t_0})$, for any given $\mu \in (0, 1)$, by taking V_{t_0} small enough. The bound in Eq. (71) depends on the initial mean state $\overline{\mathbf{s}}_{t_0}$. Also, there is no general upper bound for the largest singular value $\sigma_{t,t_0}(\overline{\mathbf{s}}_{t_0})$ itself.

Now consider the behavior of solutions of the second-moment closure equations with the nonlinear coupling term retained, so that the fundamental matrix is fully a function of $\mathbf{P}_{t_0}$, $\mathbf{M}_{t,t_0} = \mathbf{M}_{t,t_0}(\overline{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0})$. The largest singular value of $\mathbf{M}_{t,t_0}$ is in general then a function of $\mathbf{P}_{t_0}$ as well, $\sigma_{t,t_0} = \sigma_{t,t_0}(\overline{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0})$, as are the corresponding right and left singular vectors. For any particular choice of $\mathbf{P}_{t_0}$, for instance the one described in the preceding paragraph, inequality (70) still holds for every symmetric positive semidefinite matrix $\mathbf{P}$. Therefore it holds for $\mathbf{P} = \mathbf{P}_{t_0}$, and it follows that

$$\frac{V_t}{V_{t_0}} = \frac{\operatorname{tr} \mathbf{M}_{t,t_0} \mathbf{P}_{t_0} \mathbf{M}_{t,t_0}^T}{\operatorname{tr} \mathbf{P}_{t_0}} \leq \sigma_{t,t_0}^2(\overline{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0}).$$

Thus σ_{t,t_0}^2 is still an upper bound for V_t/V_{t_0}, but the property that it is necessarily attained for some choice of $\mathbf{P}_{t_0}$ has been lost.

What is actually desired is the *least* upper bound for V_t/V_{t_0}, call it $\widehat{\sigma}^2_{t,t_0}(\overline{\mathbf{s}}_{t_0})$:

$$\widehat{\sigma}^2_{t,t_0}(\overline{\mathbf{s}}_{t_0}) = \sup_{\mathbf{P}_{t_0}} \frac{\operatorname{tr} \mathbf{M}_{t,t_0}(\overline{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0}) \mathbf{P}_{t_0} \mathbf{M}^T_{t,t_0}(\overline{\mathbf{s}}_{t_0}, \mathbf{P}_{t_0})}{\operatorname{tr} \mathbf{P}_{t_0}}, \tag{72}$$

where the supremum is taken over all initial covariance matrices $\mathbf{P}_{t_0} \in \mathcal{P}_\mu(\overline{\mathbf{s}}_{t_0})$, for some given $\mu \in (0,1)$, or over all $\mathbf{P}_{t_0}$ in some chosen subset of $\mathcal{P}_\mu(\overline{\mathbf{s}}_{t_0})$. It will be convenient to take the supremum over all initial covariance matrices in the open set $\mathcal{P}_\mu(\overline{\mathbf{s}}_{t_0}; V_{\min}, V_{\max})$ for some given μ, $V_{\min}$ and $V_{\max}$, where this set is defined for $0 < \mu < 1$ and $0 \leq V_{\min} < V_{\max} \leq 2(E_{\max} - E_{\min})$ by

$$\mathcal{P}_\mu(\overline{\mathbf{s}}_{t_0}; V_{\min}, V_{\max}) = \{\mathbf{P} \in \mathcal{P}_\mu(\overline{\mathbf{s}}_{t_0}) : V_{\min} < \operatorname{tr} \mathbf{P} < V_{\max}\}. \tag{73}$$

This set is never empty, since $\mathcal{P}_\mu(\overline{\mathbf{s}}_{t_0})$ is an open set. Taking the supremum over all $\mathbf{P}_{t_0} \in \mathcal{P}_\mu(\overline{\mathbf{s}}_{t_0}; V_{\min}, V_{\max})$ makes $\widehat{\sigma}_{t,t_0}(\overline{\mathbf{s}}_{t_0})$ defined in Eq. (72) depend also on $V_{\min}$ and $V_{\max}$, that is, $\widehat{\sigma}_{t,t_0}(\overline{\mathbf{s}}_{t_0}) = \widehat{\sigma}_{t,t_0}(\overline{\mathbf{s}}_{t_0}; V_{\min}, V_{\max})$. Then one has

$$V_t/V_{t_0} \leq \widehat{\sigma}^2_{t,t_0}(\overline{\mathbf{s}}_{t_0}; V_{\min}, V_{\max}),$$

with equality either holding, or holding arbitrarily closely, for some $\mathbf{P}_{t_0} \in \mathcal{P}_\mu(\overline{\mathbf{s}}_{t_0}; V_{\min}, V_{\max})$. One can also define

$$\widehat{\sigma}_{\overline{\mathcal{T}}}(\overline{\mathbf{s}}_{t_0}; V_{\min}, V_{\max}) = \sup_{t \in \overline{\mathcal{T}}} \widehat{\sigma}_{t,t_0}(\overline{\mathbf{s}}_{t_0}; V_{\min}, V_{\max}),$$

where $\overline{\mathcal{T}} = [t_0, \overline{T}] \subseteq \mathcal{T}$ is an interval of existence of solutions for all $\mathbf{P}_{t_0} \in \mathcal{P}_\mu(\overline{\mathbf{s}}_{t_0}; V_{\min}, V_{\max})$. This yields

$$V_t/V_{t_0} \leq \widehat{\sigma}^2_{\overline{\mathcal{T}}}(\overline{\mathbf{s}}_{t_0}; V_{\min}, V_{\max}), \tag{74}$$

for all $t \in \overline{\mathcal{T}}$, with equality either holding, or holding arbitrarily closely, at some time $t \in \overline{\mathcal{T}}$ and for some $\mathbf{P}_{t_0} \in \mathcal{P}_\mu(\overline{\mathbf{s}}_{t_0}; V_{\min}, V_{\max})$.

The maximization problem of Eq. (72) is nonlinear, and in particular one cannot expect its solution to be independent of V_{t_0} as it was seen to be in the linear case, when the nonlinear coupling term in the mean equation is neglected. It is for this reason that the supremum is to be taken over the set defined in Eq. (73), which allows a range $V_{t_0} \in (V_{\min}, V_{\max})$ for the initial total variance. Inequalities (21) and (22) easily imply upper bounds for $\widehat{\sigma}^2_{\overline{\mathcal{T}}}(\overline{\mathbf{s}}_{t_0}; V_{\min}, V_{\max})$, whose dependence on $\overline{\mathbf{s}}_{t_0}$, $V_{\min}$ and $V_{\max}$ are explicit, as will be seen next.

Consider first some cases in which $\mathbf{0} \notin \mathcal{S}$, so that inequality (22) holds. Rewriting it for V_t/V_{t_0} gives

$$\frac{V_t}{V_{t_0}} < 1 + \frac{\overline{\mathbf{s}}^T_{t_0} \overline{\mathbf{s}}_{t_0} - 2E_{\min}}{V_{t_0}}. \tag{75}$$

Recall that $\overline{\mathbf{s}}_{t_0}$ and V_{t_0} must also satisfy inequality (36), rewritten here as

$$V_{t_0} < 2E_{\max} - \overline{\mathbf{s}}^T_{t_0} \overline{\mathbf{s}}_{t_0}. \tag{76}$$

The most unpredictable (largest possible V_t/V_{t_0}) case occurs when $\overline{\mathbf{s}}_{t_0}$ is near the outer boundary of $\mathcal{S}$, where the total energy is $E_{\max}$, for then V_{t_0} in inequality (75) must be small, according to inequality (76). Therefore let $\overline{\mathbf{s}}_{t_0}^T\overline{\mathbf{s}}_{t_0} = 2E_{\max} - \epsilon$, with $\epsilon > 0$ fixed and small. Then inequality (76) holds if $V_{t_0} < \epsilon$, so take the supremum over all $\mathbf{P}_{t_0} \in \mathcal{P}_\mu(\overline{\mathbf{s}}_{t_0}; \alpha\epsilon, \epsilon)$, for some given $\mu \in (0,1)$ and $\alpha \in (0,1)$. Setting $\overline{\mathbf{s}}_{t_0}^T\overline{\mathbf{s}}_{t_0} = 2E_{\max} - \epsilon$ and $V_{t_0} > \alpha\epsilon$ in inequality (75), and using inequality (74), then gives

$$\frac{V_t}{V_{t_0}} \leq \widehat{\sigma}^2_{\overline{\mathcal{T}}}(\overline{\mathbf{s}}_{t_0}; \alpha\epsilon, \epsilon) < 1 - \frac{1}{\alpha} + \frac{2(E_{\max} - E_{\min})}{\alpha\epsilon}, \tag{77}$$

for all $t \in \overline{\mathcal{T}}$. This shows that V_t/V_{t_0} can be large if ϵ is small compared to $2(E_{\max} - E_{\min})$, even if α is taken to be close to one, that is, even if one allows only a small range $V_{t_0} \in (\alpha\epsilon, \epsilon)$ for the initial total variance. The upper bound (77) for the supremum $\widehat{\sigma}^2_{\overline{\mathcal{T}}}(\overline{\mathbf{s}}_{t_0}; \alpha\epsilon, \epsilon)$ depends on the initial mean state $\overline{\mathbf{s}}_{t_0}$ only through its total energy $E_{\max} - \epsilon/2$. Recall from the derivation of inequality (22) in Sec. 4 that the bound (75) is tight if the mean state $\overline{\mathbf{s}}_t$ corresponding to $\overline{\mathbf{s}}_{t_0}$ has energy near the minimum energy level $E_{\min}$. Therefore the bound (77) for $\widehat{\sigma}^2_{\overline{\mathcal{T}}}(\overline{\mathbf{s}}_{t_0}; \alpha\epsilon, \epsilon)$ is tight if some mean state with large initial energy $E_{\max} - \epsilon/2$ approaches the minimum energy level $E_{\min}$ at any time $t \in \overline{\mathcal{T}}$.

The most predictable (smallest possible V_t/V_{t_0}) case occurs when $\overline{\mathbf{s}}_{t_0}$ is near the inner boundary of $\mathcal{S}$, where the total energy is $E_{\min}$, for then inequality (76) implies that V_{t_0} need not be small. Let $\overline{\mathbf{s}}_{t_0}^T\overline{\mathbf{s}}_{t_0} = 2E_{\min} + \epsilon$, with $\epsilon > 0$ fixed and small. Then inequality (76) holds if $V_{t_0} < V_{\max} = 2(E_{\max} - E_{\min}) - \epsilon$, so take the supremum over all $\mathbf{P}_{t_0} \in \mathcal{P}_\mu(\overline{\mathbf{s}}_{t_0}; \alpha V_{\max}, V_{\max})$, for some given $\mu \in (0,1)$ and $\alpha \in (0,1)$. Substitution into inequality (75) then gives

$$\frac{V_t}{V_{t_0}} \leq \widehat{\sigma}^2_{\overline{\mathcal{T}}}(\overline{\mathbf{s}}_{t_0}; \alpha V_{\max}, V_{\max}) < \frac{2(E_{\max} - E_{\min}) + (\frac{1}{\alpha} - 1)\epsilon}{2(E_{\max} - E_{\min}) - \epsilon}, \tag{78}$$

for all $t \in \overline{\mathcal{T}}$. This shows that V_t/V_{t_0} can never be much larger than one if ϵ is small compared to $2(E_{\max} - E_{\min})$, unless α is also taken to be small, much smaller than $\epsilon/2(E_{\max} - E_{\min})$. The explanation for this is simply that $V_t < 2(E_{\max} - E_{\min})$ for all $t \in \overline{\mathcal{T}}$ by Assumption 4, and so if V_{t_0} is already close to the value $2(E_{\max} - E_{\min})$, then there is no room for growth of the relative uncertainty V_t/V_{t_0}. This behavior is far different than what occurs when the nonlinear coupling term is neglected, for then the initial total variance V_{t_0} simply scales out of the problem.

It is expected that in many applications $E_{\min}$ and $E_{\max}$ are both known reasonably well, with $E_{\max}/E_{\min}$ not much larger than one. This still leaves plenty of room for growth of the relative uncertainty. Suppose the total energy of the initial mean state is the average of the minimum and maximum energy levels, so that $\overline{\mathbf{s}}_{t_0}^T\overline{\mathbf{s}}_{t_0} = E_{\min} + E_{\max}$. Then inequality (76) holds if $V_{t_0} < V_{\max} = E_{\max} - E_{\min}$. Again taking the supremum over all $\mathbf{P}_{t_0} \in \mathcal{P}_\mu(\overline{\mathbf{s}}_{t_0}; \alpha V_{\max}, V_{\max})$, for some given $\mu \in (0,1)$ and $\alpha \in (0,1)$, then gives simply

$$\frac{V_t}{V_{t_0}} \leq \widehat{\sigma}^2_{\overline{\mathcal{T}}}(\overline{\mathbf{s}}_{t_0}; \alpha V_{\max}, V_{\max}) < 1 + \frac{1}{\alpha}, \tag{79}$$

for all $t \in \overline{\mathcal{T}}$. If the value of α is taken to be modest, say $\alpha = 1/3$, then little growth of uncertainty can occur. However, if V_{t_0} is allowed to be small, say with $\alpha = 1/100$, then considerable growth of uncertainty can occur, relative to this small value.

In case $\mathbf{0} \in \mathcal{S}$, the strict inequality (75) simply becomes non-strict, with $E_{\min} = 0$ also. The bounds in inequalities (77)–(79) remain strict, however, because they were obtained from the strict inequality (76). Thus all that is necessary is to set $E_{\min} = 0$ everywhere following inequality (75). To see that the discussion from Eq. (72) onwards applies equally well to the stochastic process defined in Sec. 4, replace the numerator in Eq. (72) by $\operatorname{tr} \mathbf{P}_t$, where $\mathbf{P}_t$ is the covariance matrix of that process, and then re-interpret the various other quantities following Eq. (72) accordingly.

10 Conclusions

The problem of finding a system of approximate evolution equations for the mean and covariance matrix of second-order stochastic processes defined by unforced, nonlinear, conservative systems of ordinary differential equations with random initial conditions has been examined in this article from the viewpoint of energetics. A brief treatment of the stochastic initial-value problem for conservative nonlinear systems of ordinary differential equations was given, and it was used to show that the mean and covariance matrix of the resulting stochastic process are dynamically linked through an energy relationship.

The second-moment closure equations are a nonlinearly coupled system of approximate evolution equations for the mean and covariance matrix of this process. An existence and uniqueness theory was given for these equations, based largely on existence and uniqueness theory for the stochastic initial-value problem. This theory was then used to show that, under appropriate hypotheses, the mean and covariance matrix whose evolution is given by the second-moment closure equations are the mean and covariance matrix of an additional, well-defined stochastic process. It was shown further that the second-moment closure equations are energetically consistent: the mean and covariance matrix whose evolution they define are dynamically linked through precisely the same energy relationship as that of the mean and covariance matrix of the original stochastic process.

Several implications followed from this energetic consistency. One is that the total variance V_t of each of the two covariance matrices, that of the original stochastic process and that of the one whose evolution is given by the second-moment closure equations, is strictly and identically bounded in time t. It was shown further that energetic consistency implies simple, identical, time-independent bounds on the ratio V_t/V_{t_0}. Also it was shown that when the original conservative system is genuinely nonlinear, as defined in this article, total variance for the second-moment closure equations may be increasing, decreasing or stationary at times.

Essential to the results of this article is that no assumption was made on the

initial probability distribution, apart from the existence of two moments. It was also argued that the normal distribution is not appropriate in general for conservative dynamics, because it requires the existence of realizations, with nonzero probability, having total energy larger than any given amount. Furthermore, for atmospheric and ocean dynamics there are mass-like and/or temperature-like state variables that are constrained by the dynamics to be positive. With this motivation, an hypothesis was introduced that requires the realizations to have bounded energy, with probability one, and appropriate state spaces were introduced to handle state variables that are bounded from below. Many of the results were illustrated with examples.

The behavior of solutions of the second-moment closure equations was contrasted with the behavior of solutions of the approximate system obtained by neglecting the nonlinear coupling term in the mean equation. This approximate system, usually derived under an assumption of normality, is at the heart of many current large-scale computational applications in atmospheric and ocean dynamics. It was shown that this approximate system is not energetically consistent, because the nonlinear coupling term is crucial for energetic consistency.

The results of this article have left a number of open questions. For instance, it will be important to establish hypotheses under which solutions of the second-moment closure equations exist over arbitrarily long time intervals, in case solutions of the nonlinear dynamical system from which they are derived also have this property. It will also be important to develop efficient computational algorithms for implementing the second-moment closure equations in large-scale oceanic and atmospheric applications. Addressing these issues successfully will require the continued strong collaboration amongst physical scientists, computational scientists and mathematicians that has been so fruitful in recent years, as this volume attests.

Acknowledgements The author would like to thank the editors, Roger Temam and Joe Tribbia, and Lulu Stader of Elsevier, for their enthusiasm and patience during the preparation of this manuscript. The generous support of NASA's Modeling, Analysis and Prediction program, managed by Don Anderson, is also gratefully acknowledged.

References

[1] BUIZZA, R.; PALMER, T.N. [1995]: The singular vector structure of the atmospheric global circulation, J. Atmos. Sci. 52, 1434-1456.

[2] CODDINGTON, E.A.; LEVINSON, N. [1955]: Theory of Ordinary Differential Equations, McGraw-Hill, New York.

[3] COURTIER, P.; TALAGRAND, O. [1987]: Variational assimilation of meteorological observations with the adjoint vorticity equation. Part II: Numerical results, Q. J. R. Meteorol. Soc. 113, 1329-1368.

[4] DOOB, J.L. [1953]: Stochastic Processes, Wiley, New York.

[5] EPSTEIN, E.S. [1969]: Stochastic dynamic prediction, Tellus 21, 739-759.

[6] EPSTEIN, E.S.; PITCHER, E.J. [1972]: Stochastic analysis of meteorological fields, J. Atmos. Sci. 29, 244-257.

[7] FLEMING, R.J. [1971a]: On stochastic dynamic prediction I. The energetics of uncertainty and the question of closure, Mon. Wea. Rev. 99, 851-872.

[8] FLEMING, R.J. [1971b]: On stochastic dynamic prediction II. Predictability and utility, Mon. Wea. Rev. 99, 927-938.

[9] JAZWINSKI, A.H. [1970]: Stochastic Processes and Filtering Theory, Academic Press, New York.

[10] LIN, S.-J.; CHAO, W.C.; SUD, Y.C.; WALKER, G.K. [1994]: A class of the van Leer-type transport schemes and its application to the moisture transport in a general circulation model, Mon. Wea. Rev. 122, 1575-1593.

[11] MOLTENI, F.; BUIZZA, R.; PALMER, T.N.; PETROLIAGIS, T. [1996]: The ECMWF ensemble prediction system: Methodology and validation, Q. J. R. Meteorol. Soc. 122, 73-119.

[12] MOORE, A.M.; KLEEMAN, R. [1997]: The singular vectors of a coupled ocean-atmosphere model of ENSO. I: Thermodynamics, energetics and error growth, Q. J. R. Meteorol. Soc. 123, 953-981.

[13] PITCHER, E.J. [1977]: Application of stochastic dynamic prediction to real data, J. Atmos. Sci. 34, 3-21.

[14] TALAGRAND, O.; COURTIER, P. [1987]: Variational assimilation of meteorological observations with the adjoint vorticity equation. Part I: Theory, Q. J. R. Meteorol. Soc. 113, 1311-1328.

[15] THÉPAUT, J.-N.; COURTIER, P.; BELAUD, G.; LEMAÎTRE, G. [1996]: Dynamical structure functions in a four-dimensional variational assimilation: A case study, Q. J. R. Meteorol. Soc. 122, 535-561.