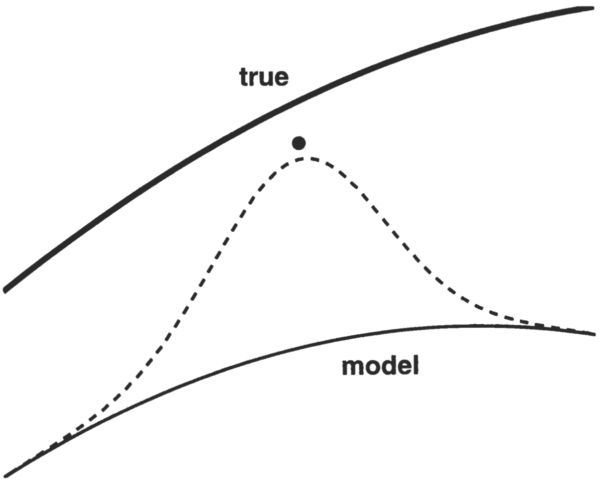

Figure 1. Sketch illustrating the possibility of data assimilation procedures overfitting the data. Heavy line depicts the true system state, the light line the forecast state, and the dashed line the analysis after data assimilation. The dot represents the observation.

will have appreciable power at small scales. If this is not altered, the analysis puts no credence in the model at small scales, so it "draws to the data" as in Figure 1. The remedy is to insist that **Q** be red or to impose a smoothness constraint in some other way. *Dee* [1991] and *Jiang and Ghil* [1993] effectively reddened **Q** by insisting that the system noise obey a geostrophic relation between pressure and velocity errors. Imposing smoothness constraints is a long-standing strategy in variational assimilation procedures [e.g., *Sasaki*, 1970]. *Provost and Salmon* [1986] dealt with the sparsity of data by restricting the analysis to a limited set of (trigonometric) basis functions. Removing small-scale variability from the basis set obviously makes local overfitting impossible. With any of these strategies, **Q** is being forced to have fewer significant degrees of freedom than its full size would allow.

The arguments presented above prompt the strategy described in the next section. For purposes of specifying **Q** and calculating the **P** we reduce the state space to the small number of degrees of freedom adequate to carry what little we know about the system noise, while maintaining enough of the KF's ability to use the dynamics to propagate error information. Inter alia, this reduction ensures a satisfactory degree of smoothness. The procedure may also be viewed as a parameterization of the large matrices **Q** and **P** in terms of a relatively small number of parameters. Our version of (5) propagates these parameters in time rather than the full model error covariance. Full covariance matrices can be reconstituted from the parameters.

3. General Method

We seek representations of the error covariance matrices with fewer degrees of freedom than implied by the dimension of the model state space. We begin by finding a reduced representation of the model state space and a reduced model to accomplish the transitions from one update time to the next. Similar developments are given by FMR, *Xue et al.* [1994], and Y. Xue et al. (Predictability of a coupled model of ENSO using singular vector analysis. Part 1: Optimal growth in seasonal background and ENSO cycles, submitted to *Monthly Weather Review*, 1996; hereinafter referred to Xue et al., submitted manuscript, 1996). Let **w** be the vector of all state space variables. For example, if the model variables are u,v,p defined at the set of space points $\mathbf{x}_u,\mathbf{x}_v,\mathbf{x}_p$, respectively, then

$$\mathbf{w}(\boldsymbol{x},t) = (\mathbf{u}(\mathbf{x}_u,t),\ \mathbf{v}(\mathbf{x}_v,t),\ \mathbf{p}(\mathbf{x}_p,t))^T.$$

There is no need for the different model variables to be defined on the same grid or with the same number of points. Now write the state space vector in terms of factors in time and space:

$$\mathbf{w}(\boldsymbol{x},t) = \mathbf{E}(x)\,\mathbf{u}(t). \tag{9}$$

The equality in (9) means that the columns of the $N \times N$ matrix **E** are a complete set of basis functions for the model state space. The vector $\mathbf{u}(t)$ holds the amplitudes at time t of these basis functions. There are many possible choices for the columns of **E**, e.g., Fourier components for each model variable. The important requirement for us is that the choice is efficient in that it lets us truncate the number of columns in **E** (and hence the dimension of **u**) to M << N without sacrificing anything essential. We choose multivariate EOFs (MEOFs) for the columns of **E**, in which case the elements of **u** are the principal components (PCs). It is convenient to take the basis set to be orthonormal; $\mathbf{E}^T\mathbf{E} = \mathbf{I}$. Because of the truncation it is not also true that $\mathbf{E}\mathbf{E}^T = \mathbf{I}$.

Then since

$$\mathbf{u} = \mathbf{E}^T\mathbf{w}, \tag{10}$$

left multiplying the model evolution equation (4) by $\mathbf{E}^T$ yields

$$\mathbf{u}^f(t+1) = \mathbf{A}\,\mathbf{u}^a(t) + \mathbf{E}^T\boldsymbol{\Phi}\,\tau(t) + \mathbf{u}' \tag{11}$$

where

$$\mathbf{A} = \mathbf{E}^T\boldsymbol{\Psi}\,\mathbf{E} \tag{12}$$

and $\mathbf{u}'$ accounts for the influence of the discarded modes at time t on the retained ones at time t+1. As in FMR, we assume that $\mathbf{u}'$ is negligible. For our sea level simulations it is easy to make this true by retaining enough MEOFs. As appears below, it is more problematic for error propagation.

There is a straightforward way to calculating the transition matrix **A** (cf. FMR or Xue et al. (submitted manuscript, 1996) for details). With **A** in hand we find the error covariances **P** in the reduced representation from reduced state space versions of (5), (6), and (2):

$$\mathbf{P}^f(t+1) = \mathbf{A}\,\mathbf{P}^a(t)\,\mathbf{A}^T + \mathbf{Q}(t); \tag{13}$$

$$\mathbf{P}^a = (\mathbf{I} - \mathbf{KH})\,\mathbf{P}^f; \tag{14}$$

$$\mathbf{K} = \mathbf{P}^f\mathbf{H}^T\,(\mathbf{H}\,\mathbf{P}^f\mathbf{H}^T + \mathbf{R})^{-1} \tag{15}$$

While these equations are formally identical to the earlier ones (apart from the change from Ψ to **A**), the meanings of the symbols have changed. Here

$$\mathbf{P}^{a,f} = \langle(\mathbf{u}^{a,f} - \mathbf{u}^{\mathrm{true}})\,(\mathbf{u}^{a,f} - \mathbf{u}^{\mathrm{true}})^T\rangle \tag{16}$$

and **Q** is the system noise appropriate to the reduced system. There is a relation between the new reduced mapping matrix **H** and the original full state space one $\mathbf{H}^\dagger$

$$\mathbf{H} = \mathbf{H}^\dagger\mathbf{E}, \tag{17}$$

which shows that the observational data are now approximated