Assessing the performance of dynamical trajectory estimates

Jochen Bröcker
School of Mathematical and Physical Sciences, University of Reading, Whiteknights, P.O. Box 220, Reading RG6 6AX, United Kingdom

(Received 4 April 2014; published 18 June 2014)

Estimating trajectories and parameters of dynamical systems from observations is a problem frequently encountered in various branches of science; geophysicists for example refer to this problem as *data assimilation*. Unlike as in estimation problems with exchangeable observations, in data assimilation the observations cannot easily be divided into separate sets for estimation and validation; this creates serious problems, since simply using the same observations for estimation and validation might result in overly optimistic performance assessments. To circumvent this problem, a result is presented which allows us to estimate this optimism, thus allowing for a more realistic performance assessment in data assimilation. The presented approach becomes particularly simple for data assimilation methods employing a linear error feedback (such as synchronization schemes, nudging, incremental 3DVAR and 4DVar, and various Kalman filter approaches). Numerical examples considering a high gain observer confirm the theory.

DOI: 10.1103/PhysRevE.89.062919 PACS number(s): 05.45.Tp, 05.45.Xt, 02.50.Fz, 92.60.Wc

I. INTRODUCTION

Generally speaking, the question to be revisited in this paper is the following: "Is my model consistent with my observed data?" We consider a situation in which the model is a dynamical system and the observations are noise corrupted. In order to meaningfully assess the model, the observations have to be compared with model *trajectories*. These trajectories though have to be estimated using the observations, even if the model itself is fully specified. This form of dynamical estimation problem will be referred to as *data assimilation*, a term borrowed from geophysical applications (see, e.g., Refs. [1–3]). It might be tempting to use the residual error of these trajectories with respect to the observations as a measure of model performance. Thereby, however, the observations would be used *both* to estimate and to evaluate the trajectories. This "in sample" evaluation might result in overly optimistic performance assessments.

As a simple example, the reader might think of trying to synchronize two systems uni–directionally over a noisy channel. If the coupling is too weak, no synchronization will occur. With increasing coupling, the error with respect to the observations (the in-sample observational error) will typically become ever smaller (see also Fig. 1, circles, for results from a numerical experiment). The error with respect to the underlying clean signal though (the out-of-sample observational error) will at some point become again *larger* as the slave system is increasingly affected by the noise (Fig. 1, diamonds). This can be considered as a form of overfitting.

In statistical regression with independent or at least exchangeable observations, this problem can be circumvented by separating the data into disjoint sets for estimation and validation, or at least by mimicking such a separation by cross validation or similar procedures [4]. These approaches though affect the temporal structure of the observations and are therefore not directly applicable to data assimilation. In this paper, an alternative is presented, also inspired by similar results from statistical regression (e.g., the C_p criterion [4]). The idea is to estimate the optimism of the in-sample observational error in data assimilation. The optimism is essentially due to correlations between the trajectory estimates and the observational noise. Obviously, these correlations are an inevitable consequence of the fact that the trajectory estimates are functions of the observations.

Our discussion requires a brief digression on the quadratic error in continuous time as well as some aspects of stochastic integration. Next, we derive the main mathematical result of this paper. In terms of applying the presented theory in practice, two general classes of data assimilation approaches are discussed separately. The first class are algorithms of the filtering type where the estimated system state at any time t is *nonanticipating*, that is, dependent only on past and present, but not future, observations. This encompasses a very large class of algorithms relevant in physics, engineering, and the atmospheric sciences, including various approximations of the optimal nonlinear filter [5] and algorithms for incremental data assimilation such as nudging and synchronization schemes, 3DVAR, and 4DVar (see Ref. [2] for an overview). In fact, all of these satisfy the even stronger assumption of linear observational error feedback. The second class are algorithms of the smoothing type where the estimated system state at any time t depends on the entire observational record, but the estimated trajectories are differentiable. This includes for example weak-constraint (long window) 4DVar [6]. We stress here that, to the best of our knowledge, all data assimilation approaches of practical relevance are of either the filtering type or the smoothing type. In this paper, we focus mainly on filtering schemes, and it turns out that the optimism of the in-sample observational error for these schemes obeys a very simple formula which can be evaluated with only a few additional computations. The results are applied in a numerical example employing synchronization. This example confirms that the observational error becomes increasingly overoptimistic with stronger coupling, while the presented theory correctly predicts the optimism (which we can compute directly in our synthetic experiment).

II. THE QUADRATIC ERROR IN CONTINUOUS TIME

In continuous time, we consider a fixed time interval $[0,T]$ and define the observations as

$$o_t = \zeta_t + \sigma r_t, \quad t \in [0,T], \tag{1}$$

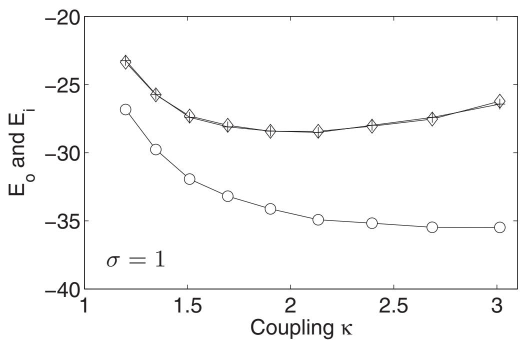

FIG. 1. Synchronization errors vs coupling strength for two Lorenz'63 systems with unidirectional coupling over the noisy channel. The in-sample error (○) decreases with increasing coupling strength, while the out-of-sample error (◇) shows a clear minimum for a finite coupling. The out-of-sample error agrees with the estimates (+) provided by the right-hand side of Eq. (15) (with sample averages replacing expectation values; see also accompanying text). Results are shown for $\sigma = 1$; other noise intensities gave qualitatively similar results.

with $\zeta_.$ being the desired signal and $r_.$ the noise. (The subscript dot notation is used to refer to an entire time series, e.g., $\zeta_.$ is a shorthand for $\{\zeta_t, t \in [0,T]\}$.) The noise $r_.$ is supposed to be a continuous, stochastic signal of bandwidth $\propto 1/\delta$ which approximates white noise with unit intensity for $\delta \to 0$. This will have to be made precise later. The desired signal $\zeta_.$, henceforth referred to as the *output*, which we assume to be continuous, carries the relevant information. We think of $\zeta_.$ as being generated through some dynamical system (deterministic or stochastic); for example, $\zeta_.$ might be of the form $\zeta_t = h(z_t)$, where h is some function and $z_.$ the solution of an ordinary differential equation with unknown initial condition, or more generally a Markov process. However, the details of how $\zeta_.$ arises are irrelevant at this point.

The result of our data assimilation procedure is a process $\{x_t, t \in [0,T]\}$, which we assume to be continuous. Further, x_t for any $t \in [0,T]$ might depend on the entire observational record $o_.$. (Note that our setup also includes the case where $x_.$ depends on model parameters which in turn have been estimated from the observations [7].) It now seems natural to define the (mean square) observational error as

$$q(x_., o_., \delta) = \frac{1}{T}\int_0^T (x_t - o_t)^2 dt; \tag{2}$$

this definition is very preliminary only, as we now want to take the limit $\delta \to 0$. If we expand q, we obtain

$$q(x_., o_., \delta) = \frac{1}{T}\int_0^T x_t^2 dt + \frac{1}{T}\int_0^T o_t^2 dt - \frac{2}{T}\int_0^T x_t o_t dt. \tag{3}$$

The limit $\delta \to 0$ should be seen as an approximation to the case of large but finite bandwidth encountered in practice and is used for convenience as explicit dependence on the bandwidth would make our results unnecessarily complicated and specific. The first term in Eq. (3) does not depend on δ. The second term though does and in fact grows like the bandwidth of the noise. On the other hand, this term does not depend on $x_.$. Hence, as long as only performance *differences* (for example, between two competing data assimilation algorithms) are of interest, a modified mean square error with this term simply being omitted still makes sense. If the third term converges, we denote the resulting random quantity by

$$\lim_{\delta \to 0} \int_0^T x_t o_t dt = \int_0^T x_s \circ d\eta_s, \tag{4}$$

with the *accumulated observations* defined as

$$\eta_t = \int_0^t \zeta_s ds + \sigma W_t. \tag{5}$$

With this definition we can write

$$\int_0^T x_s \circ d\eta_s = \int_0^T x_s \zeta_s ds + \sigma \int_0^T x_s \circ dW_s,$$

where $W_.$ is the standard Wiener process [12]. The first term on the right-hand side is a standard Riemann integral, while the interpretation (and existence) of the second *stochastic integral* requires some discussion which will be provided presently. Our definite mean square error (denoted with Q) now reads as

$$Q(x_., \eta_.) = \frac{1}{T}\int_0^T x_s^2 ds - \frac{2}{T}\int_0^T x_s \circ d\eta_s, \tag{6}$$

provided the stochastic integral makes sense.

For the remainder of this section, we briefly mention relevant properties of stochastic integrals (further material has been included in Appendix B; see, e.g., Refs. [13,14] for more information). The stochastic integral $\int_0^t x_s \circ dW_s$ is defined if one of the following two sets of assumptions apply (both are relevant for data assimilation).

A. Smooth $x_.$

If x_t is a smooth (by which we mean differentiable) function of t, then the stochastic integral is well defined and independent from the details of the approximation of the white noise. It is then permitted to integrate by parts

$$\int_0^T x_s \circ dW_s = x_T W_T - \int_0^T \dot{x}_s W_s ds, \tag{7}$$

whereby we have expressed the stochastic integral through a standard Riemann integral (note that W_s is a continuous function of s).

B. Nonanticipating $x_.$

The stochastic integral can still be defined if $x_.$ is continuous and nonanticipating; that is, for any t, the random variable x_t is independent of the future increments $\{W_{t+s} - W_t, s \leqslant 0\}$. The value of the integral however depends to some extent on how the white noise is approximated by limited bandwidth noise. This is not an artifact of the continuous time limit but a manifestation of phenomena already present in discrete time. Here we use the presumably very natural approximation with piecewise linear functions; the details have been deferred to Appendix A. The integral is then referred to as the *Stratonovic integral*.

We stress that the two definitions in Secs. A and B are consistent in that if $x_.$ happens to be smooth *and* nonanticipating, any approximation to the Stratonovic integral will in the limit satisfy Eq. (7). Finally, if $x_.$ and $W_.$ are independent, then in both cases A and B we have $\mathbb{E}[\int_0^t x_s \circ dW_s] = 0$.

III. THE OUT-OF-SAMPLE ERROR

The quadratic error $Q(x_.,\eta_.)$ between output and observations is what we compute in practice, but what we should really be interested in is the error $Q(x_.,\zeta_.)$ between output and desired signal. Although the latter cannot be computed directly, an interesting and useful relation between the two can be derived. By straightforward calculation using the definitions of Q and η, we obtain

$$Q(x_.,\zeta_.) = Q(x_.,\eta_.) + \frac{2\sigma}{T}\int_0^T x_s \circ dW_s. \tag{8}$$

Defining the *in-sample* and *out-of–sample* errors as $E_I = \mathbb{E}[Q(x_.,\eta_.)]$ and $E_O = \mathbb{E}[Q(x_.,\zeta_.)]$, respectively, we can take the expectation of Eq. (8) and obtain

$$E_O = E_I + \frac{2\sigma}{T}\mathbb{E}\left[\int_0^T x_s \circ dW_s\right]. \tag{9}$$

This identity (which should be regarded as the "dynamical version" of Eq. (7.22) in Ref. [4]) is the basis of our further analysis. The second term in Eq. (9) is referred to as the *optimism* of the in-sample error. The optimism is essentially a correlation between $x_.$ and $W_.$. This correlation is an inevitable consequence of the fact that $x_.$ is correlated with the observations. We see that although *some* correlation between output and observations is required in order that $x_.$ captures some information from the observations, *too much* correlation entails that the in-sample error E_I is a (typically upward) biased estimate of the out-of-sample error E_O.

These considerations show that data assimilation (or in fact any estimation procedure employing noisy observations) requires one to carefully tune the amount of correlation between observations and estimated trajectories. Here is where Eq. (9) becomes practically relevant. The optimism depends largely on the employed data assimilation approach and *not* on the details of the underlying "true dynamics." Most importantly though, Eq. (9) (or its approximations) allows one to calculate the out-of-sample error *without* the need of independent test observations.

Typically, approximations are necessary in order to apply Eq. (9) in practice. First, one has to make do with taking the empirical average in-sample-error $Q(x_.,\eta_.)$ instead of its expectation E_I (unless the latter can be computed explicitly). Second, calculating $\mathbb{E}[\int_0^T x_s \circ dW_s]$ requires approximations if $x_.$ is a complicated function of the observations. Further, the standard deviation σ is required, which essentially amounts to saying what part of the observations $\eta_.$ is desired and which not. This is an approximation but in the same sense that noise itself is an approximation. Most importantly, the noise level σ of a measurement device can often be estimated from a separate experiment (or gathered from the manufacturer's instructions), unlike the optimism of a specific trajectory estimate. We also stress that the formalism cannot be applied if there is multiplicative noise in the observations, i.e., if σ is random.

To further analyze the optimism and obtain more specific results, two classes of data assimilation schemes are now considered, namely, of the smoothing type and the filtering type. Our discussion of smoothing type approaches is brief, because a very detailed analysis of this situation in the context of the minimum energy estimator has been carried out in Ref. [15]. We then focus on filtering type approaches.

A. Smoothing type approaches

Various data assimilation approaches exist which result in estimates that are anticipating; i.e., the estimate x_t at any time t depends on the entire observational record $\eta_.$. The optimism of the in-sample error is still well defined [via Eq. (7)] if $x_.$ is smooth, which we take to mean "differentiable" here. Examples are the maximum *a posteriori* estimator [16] and the minimum energy estimator [17,18], which can be seen as a continuous time analog of weakly constrained 4DVAR [19]. A detailed analysis of the presented theory in the case of the minimum energy estimator has been carried out in Ref. [15]. The essence of that research is that the optimism is strongly related to the sensitivity of the estimated trajectory with respect to the observations. To very good approximation, the sensitivity can be obtained from the Hessian of the cost function.

Here, we discuss a much simpler but very illustrative application to orthogonal function expansions. Consider differentiable functions $\phi_m : [0,T] \to \mathbb{R}$ for $m = 1,\dots,M$ with the property $\int_0^T \phi_m(t)\phi_n(t)dt = \delta_{mn}$. We set $x_t = \sum_m a_m\phi_m(t)$ and determine the coefficients a_m by minimizing the mean square error. Using the orthonormality of ϕ_m we obtain

$$Q(x_.,\eta_.) = \sum_m a_m^2 - 2\sum_m a_m \int_0^T \phi_m(t)d\eta_t,$$

which is minimal for $a_m = \int_0^T \phi_m(t)d\eta_t$. With a_m being set thus and using elementary properties of stochastic integrals, we obtain $\mathbb{E}[\int_0^T x_t \circ dW_t] = \sigma M$, where M is the number of free parameters in the model. Equation (9) then gives $E_O = E_I + \frac{2\sigma^2 M}{T}$. This formula is the continuous time equivalent of the famous C_p criterion from linear regression, which is obtained if T is interpreted as "the number of samples." It is intuitively clear that using an increasing number M of basis functions will cause E_I to decrease. More specifically, $E_I = -C_M - \frac{\sigma^2 M}{T}$, where C_M is positive increasing and converges to $\int_0^T \zeta_t^2 dt$, the energy of the desired signal. Hence $E_O = -C_M + \frac{\sigma^2 M}{T}$ must have a minimum at some finite M, which can be interpreted as the optimal number of basis functions for reconstructing the desired signal.

B. Filtering type approaches

We now consider data assimilation approaches giving nonanticipating estimates; i.e., the estimate x_t at any time t depends only one the observational record η_s, $s \leqslant t$. Any smoothing type approach yields also filtering type estimates; indeed, if we consider a smoothing type estimate over the reduced observational history η_s, $s \in [0,\tau]$ with $\tau < T$, then the endpoint x_τ of the estimated trajectory is nonanticipating. By incrementing τ as new observations come in, a filtering type algorithm is obtained. It turns out that filtering type algorithms thus obtained employ a *linear error feedback*. This entails that

$x_.$ obeys a stochastic differential of the form

$$dx_t = \alpha_t dt + b_t d\eta_t, \tag{10}$$

which can be written as $dx_t = a_t dt + \sigma b_t dW_t$, with $a_t = \alpha_t + b_t \zeta_t$. These equations are to be understood in the Ito sense; $a_.$ and $b_.$ are assumed to be nonanticipating, whence $x_.$ is also nonanticipating. This is the case, for example, if for all t, a_t and b_t depend on the history of observations up to and including t, plus maybe further random components which are entirely independent of the observational record, for example, initial conditions. Apart from the filtering type algorithms mentioned above, these assumptions in fact cover the optimal nonlinear filter if either the mode or the first moment is taken as the estimate (see Ref. [5], Eq. (6.116) or (6.95), respectively), and also its various approximations such as the extended Kalman filter and continuous time versions of the ensemble Kalman filter (as discussed in Ref. [20]), as well as nudging and synchronization approaches [21].

The core step towards computing the optimism for filtering type algorithms is to convert the Stratonovic integral in Eq. (9) to a sum of Riemann and Ito integrals. Using stochastic calculus (more specifically Ref. [13], Chap. 3, Def. 3.13) and Eq. (10), it is easy to see that

$$\int_0^t x_s \circ dW_s = \int_0^t x_s dW_s + \frac{\sigma}{2}\int_0^t b_t dt, \tag{11}$$

and since the expectation value of the Ito integral is zero, combining Eqs. (9) and (11) gives

$$E_O = E_I + \frac{\sigma^2}{T}\int_0^T \mathbb{E}[b_t]dt. \tag{12}$$

The computation of the optimism should not represent any difficulties whatsoever if b_t is purely a function of time (as is the case in Kalman filtering and synchronisation/nudging). There exist data assimilation schemes of filtering type where b_t depends implicitly on the observations and is therefore random (as, e.g., in continuous time variants of the extended Kalman filter). We might still get reasonable approximations by assuming

$$\frac{\sigma^2}{T}\int_0^T b_t dt \cong \frac{\sigma^2}{T}\int_0^T \mathbb{E}[b_t]dt. \tag{13}$$

In the limit $T \to \infty$, this might be rigorously true due to appropriate ergodic properties of the problem.

IV. NUMERICAL EXAMPLE: SYNCHRONIZATION

We discuss a numerical experiment where $x_.$ is obtained using a certain synchronization scheme. In this experiment, b_t in Eqs. (10) and (12) will simply be a constant. Our choice of b_t is motivated by the high gain observer concept [23]. Assume for the moment that there is no noise and suppose that $\zeta_t = z_t^{(1)}$, the first component of some $z_t \in \mathbb{R}^d$, which in turn satisfies a differential equation of the form

$$\dot{z}_t = f(z_t), \quad t \in [0,T].$$

The *observer* is the dynamical system

$$\dot{\xi}_t = f(\xi_t) + K(x_t - \zeta_t), \quad x_t = \xi_t^{(1)}, \tag{14}$$

where the gain K is a $d \times 1$ matrix. We have $d = 3$ in our example. The error $e_t = x_t - \zeta_t$ can be made to decrease exponentially fast by ensuring that the roots of the polynomial $\chi(\lambda) = \lambda^3 - K^{(1)}\lambda^2 - K^{(2)}\lambda - K^{(3)}$ are sufficiently far in the left half of the complex plane. (Strictly speaking, for this result to apply, the system has to evolve on a compact set.) Hence if we choose $\chi(\lambda) = (\lambda + \kappa)^3$, and the gain K_κ accordingly, the high gain observer theory predicts that synchronization will occur if κ is large enough. We refer to κ as the coupling parameter.

In the presence of noise, the observer (14) modifies to

$$d\xi_t = [f(\xi_t) + K_\kappa x_t]dt - K_\kappa d\eta_t, \quad x_t = \xi_t^{(1)}.$$

Comparison with Eq. (10) yields that now $b_t = -K_\kappa^{(1)} = 3\kappa$; hence Eq. (12) gives the simple formula

$$E_O = E_I + 3\sigma^2\kappa. \tag{15}$$

Note that the optimism only depends on the coupling but not on the actual dynamics. We will now verify this statement numerically. In order to do this, we replace E_I and E_O in Eq. (15) with their sample approximations $Q(x_.,\eta_.)$ and $Q(x_.,\zeta_.)$, respectively. (The uncertainty introduced through this approximation will be assessed through Monte Carlo resampling.) We further compute $3\sigma^2\kappa$ and check if these three quantities satisfy Eq. (15). The value of Eq. (15) is in that it would allow us to express $Q(x_.,\zeta_.)$, which is inaccessible in realistic situations, through $Q(x_.,\zeta_.) + 3\sigma^2\kappa$, which is accessible.

For the vector field f we use the chaotic Lorenz'63 system given by

$$f(x,y,z) = [S(y-x), Rx - y - xz, xy - Bz]^t,$$

with parameters $S = 10$, $R = 28$, and $B = 8/3$. For the observer, perturbed parameters $S = 9.9$, $R = 27.2$, and $B = 2.63$ were used. For this setup, we found the synchronization threshold to be at $\kappa \cong 1.2$.

A total of 99 simulations were performed for noise intensities σ between 0.25 and 4 and coupling parameters κ between 1.2 and 3. A stochastic Euler scheme of fixed time step $\Delta t = .005$ for a time interval with $T = 250$ was used (an initial transient was discarded); this corresponds to 50 000 sample points. Each simulation was repeated 20 times with different realizations of the noise.

The simulations yielded a number of interesting (yet maybe not surprising) facts. Figure 1 shows the results for $\sigma = 1$. The coupling parameter varies along the abscissa. Other noise intensities gave qualitatively similar results. First, the in-sample error (marked with circles) decreases with increasing coupling strength, while the out-of-sample error (marked with diamonds) shows a clear minimum. Second, the results appear to be consistent with Eq. (15) (the out-of-sample error estimates are marked with crosses, +). The last point was investigated further. In Fig. 2, the optimism, that is, the differences between in-sample and out-of-sample errors are shown for three different values of σ. From the 20 independent simulations, ± 2 standard deviation consistency bars were calculated, providing an indication of the variability of the optimism across different simulations. The straight line gives the optimism according to Eq. (15). All quantities have been

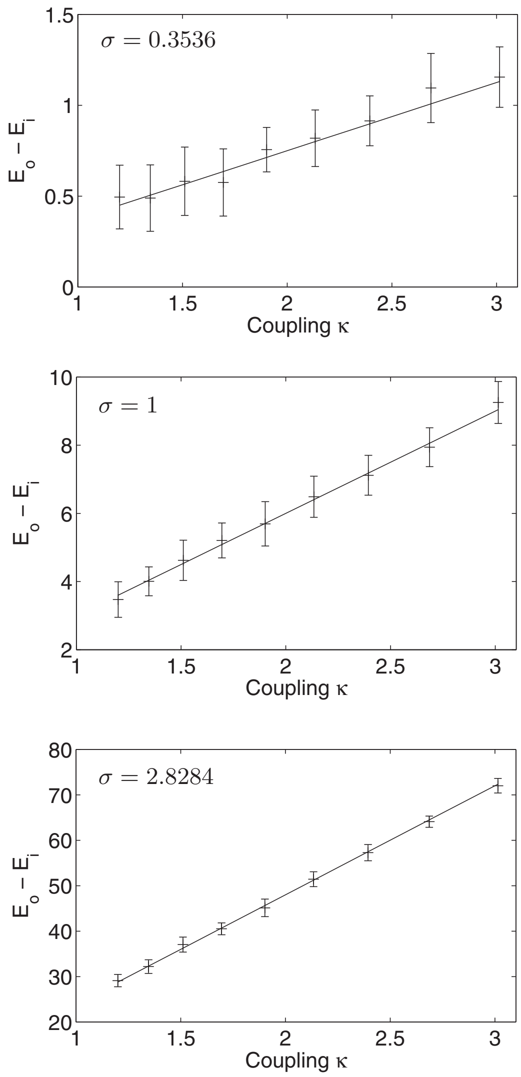

FIG. 2. Difference between in-sample and out-of-sample errors (+ with error bars) for several noise levels. Error bars were obtained from repeating the experiment with independent noise realizations. Solid line shows the same quantity as estimated from our theory (i.e., $3\sigma^2\kappa$, see right-hand side of Eq. (15) and accompanying text).

normalized with the variance of the signal $\zeta_.$. It emerges that the theoretical estimate agrees very well with the empirical optimism, confirming our theory qualitatively and quantitatively. It is also evident that the optimism grows with increasing coupling strength. Very similar experiments were conducted using Chua's circuit; the findings were exactly the same.

V. CONCLUSIONS

When estimating trajectories of a dynamical system from observations, the error with respect to the observations is often a too optimistic estimator of performance, since the observations have been used already to find the trajectory estimate. This optimism was investigated here in a situation where observations obtain continuously in time and are corrupted with additive white noise. As a measure of deviation, we have considered the mean square error. An estimate of the optimism was presented which depends only on quantities which are available in an operational situation. This result has potential application in the assessment of data assimilation techniques. A detailed numerical simulation was presented for the high gain observer, where the presented theory could be used to optimize the feedback gain (aka the coupling constant).

ACKNOWLEDGMENTS

Fruitful discussions with Martin Kolb are gratefully acknowledged. Ian Melbourne reminded me of the fact that the results of Wong and Zakai on the smooth approximation of stochastic integrals do not necessarily extend to higher dimensions. Careful comments of several anonymous referees further helped to improve the manuscript.

APPENDIX A: PIECEWISE LINEAR APPROXIMATION

We assume the situation in Sec. II B, i.e., $x_.$ is nonanticipating and $W_.$ is the standard Wiender process on the interval $[0,T]$. For the partition $0 = t_1 < \cdots < t_N = T$, $t_{n+1} - t_n \leqslant \Delta t$ of that interval we define

$$W_t^{\Delta t} = \frac{t - t_n}{t_{n+1} - t_n} W_{t_{n+1}} + \frac{t_{n+1} - t}{t_{n+1} - t_n} W_{t_{n+1}}$$

for $t_n \leqslant t < t_{n+1}$. Clearly, this is a piecewise linear approximation of $W_.$ which collocates at the points of the partition. We now define

$$\int_0^T x_s \circ dW_s = \lim_{\Delta t \to 0} \int_0^T x_s \dot{W}_s^{\Delta t} ds$$

provided the limit exists.

APPENDIX B: STOCHASTIC INTEGRALS

The stochastic integral $\int_0^T x_s \circ dW_s$ discussed in Sec. II is known as a Stratonovic integral. However, it is commonly defined as the limit $\Delta t \to 0$ of

$$\sum_{n=1}^{N} \frac{x_{t_n} + x_{t_{n+1}}}{2} (W_{t_{n+1}} - W_{t_n}), \tag{B1}$$

where $\{t_n\}$ is a partition as in Appendix A. It follows from results in Ref. [14], Chap. 6, Sec. 7, that this gives the same result as our previous definition in Appendix A. The way the process $x_.$ enters in Eq. (B1) is crucial for the definition of the Stratonovic integral. Taking for example $x_{t_n}(W_{t_{n+1}} - W_{t_n})$ instead as summands in definition (B1) gives another integral referred to as the Ito integral. Both Ito and Stratonovic integrals differ in general. It is customary in mathematics to write Stratonovic and Ito integrals as $\int_0^T x_s \circ dW_s$ and $\int_0^T x_s dW_s$, respectively. A very important property of the Ito integral is that $\mathbb{E}[\int_0^t x_s dW_s] = 0$ for any t. This is not true for the Stratonovic integral.

[1] K. Ide, P. Courtier, M. Ghil, and A. C. Lorenc, J. Meteorol. Soc. Jpn. **75**, 181 (1997).
[2] E. Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability*, 1st ed. (Cambridge University Press, Cambridge, UK, 2001).
[3] G. Evensen, *Data Assimilation: The Ensemble Kalman Filter* (Springer-Verlag, New York, 2007).
[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. (Springer, New York, 2009), p. 524.
[5] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*, Mathematics in Science and Engineering, Vol. 64 (Academic Press, New York, 1970).
[6] M. Fisher, Y. Trémolet, H. Auvinen, D. Tan, and P. Poli, *Weak Constraint and Long Window 4D-Var*, Techical Report 655 (European Centre for Medium Range Weather Forecasting, Reading, United Kingdom, 2011).
[7] The literature on identifying dynamical models is of course vast: see, e.g., Ref. [8] for methods for general stochastic differential equations; Ref. [9] for approaches from nonlinear time series analysis; or the classics, Refs. [10,11], for the linear theory. Clearly this list is very far from being complete.
[8] M. Kessler, A. Lindner, and M. Sørensen (eds.), *Statistical Methods for Stochastic Differential Equations*, Monographs on Statistics and Applied Probability, Vol. 124 (CRC Press, Boca Raton, FL, 2012), revised papers from the 7th Séminaire Européen de Statistique on Statistics for Stochastic Differential Equations Models, Cartagena, May 7–12, 2007.
[9] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, UK, 1997).
[10] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods* (Springer-Verlag, New York, 1987).
[11] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, 3rd ed. (Prentice Hall, New York, 1994).
[12] Aka standard Brownian motion or "integrated white noise" with unit intensity.
[13] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus*, 2nd ed., Graduate Texts in Mathematics, Vol. 113 (Springer-Verlag, New York, 1991).
[14] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, 2nd ed., North-Holland Mathematical Library, Vol. 24 (North-Holland, Amsterdam, 1989), pp. xvi+555.
[15] J. Bröcker and I. G. Szendro, Q. J. R. Meteorol. Soc. **138**, 785 (2012).
[16] O. Zeitouni and A. Dembo, Stochastics **20**, 221 (1987).
[17] R. E. Mortensen, J. Optim. Theory Appl. **2**, 386 (1968).
[18] O. Hijab, Ann. Probab. **12**, 890 (1984).
[19] Y. Sasaki, Mon. Weather Rev. **98**, 875 (1970).
[20] L. Lei, D. R. Stauffer, S. E. Haupt, and G. S. Young, Tellus, Ser. A **64** (2012), doi:10.3402/tellusa.v64i0.18484.
[21] It is worth stressing that, in control theory, there exists an extensive body of literature on deterministic observers employing *nonlinear* feedback (see, e.g., Ref. [22]). These observers though cannot readily be generalized to accommodate noisy observations.
[22] H. Nijmeijer and A. van der Schaft, *Nonlinear Control Systems* (Springer, New York, 1990).
[23] H. Nijmeijer and I. M. Mareels, IEEE Trans. Circuits Syst, I: Fundam. Theory Appl. **44**, 882 (1997).