PERGAMON

Automatica 37 (2001) 325–340

automatica

www.elsevier.com/locate/automatica

Computationally efficient steady-state multiscale estimation for 1-D diffusion processes☆

Terrence T. Ho[a], Paul W. Fieguth[b,*], Alan S. Willsky[c]

[a]*Booz Allen & Hamilton, New York, USA*
[b]*Systems Design Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1*
[c]*Massachusetts Institute of Technology, Cambridge, MA, USA*

Received 15 April 1998; revised 20 September 1999; received in final form 4 July 2000

Abstract

Conventional optimal estimation algorithms for distributed parameter systems have been limited due to their computational complexity. In this paper, we consider an alternative modeling framework recently developed for large-scale static estimation problems and extend this methodology to dynamic estimation. Rather than propagate estimation error statistics in conventional recursive estimation algorithms, we propagate a more compact multiscale model for the errors. In the context of 1-D diffusion which we use to illustrate the development of our algorithm, for a discrete-space process of N points the resulting multiscale estimator achieves $\mathcal{O}(N \log N)$ computational complexity (per time step) with near-optimal performance as compared to the $\mathcal{O}(N^3)$ complexity of the standard Kalman filter.

Keywords: Multiscale realization; Dynamic estimation; Distributed parameter systems; Diffusion

1. Introduction

Estimation for distributed parameter systems governed by partial differential equations (PDEs), such as those found in applications ranging from pollution control (Omatu, Seifeld, Soeda, & Sawaragi, 1988) to the modeling of ecological systems and flexible structures (Banks & Kunisch, 1989), has received considerable attention in the estimation and control communities in the past. While there have been successful applications of the theory of optimal estimation for such systems, it is also true that there are severe computational barriers that limit the domain in which truly optimal methods can be implemented. Indeed this is certainly the case in the field of remote sensing in which "data assimilation", the melding of data with dynamic models, represents one of the most significant current-day problems. For example, in problems of atmospheric or oceanographic data assimilation (Fieguth, Karl, Willsky, & Wunsch, 1995), the dimensionality of finite-dimensional approximations to the underlying dynamics can range from hundreds of thousands to hundreds of *millions*. Given the need in such applications to produce both estimates and estimation error variances, the computational challenge is substantial. Indeed, conventional linear least-squares estimation (LLSE) algorithms, such as Kalman filtering, are completely impractical for solving such large problems both for computational and for storage reasons. A critical aspect of these estimation problems is the requirement that estimation error statistics be computed. This necessity precludes the use of accelerated methods such as multigrid (Briggs, 1987), which do not supply such error statistics, or the FFT, which requires spatially stationary prior models and spatially regular measurement patterns, requirements that cannot be met in many applications including most, if not all, remote sensing problems.

For these reasons, it is clear that there is a need for suboptimal (that is, approximate) estimation algorithms that can deal effectively with the computational challenges. The key to doing this is to find a compact and effective representation for the statistics of the estimation

☆This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised form by Associate Editor A. Bagchi under the direction of Editor Tamer Basar. This work was supported by ONR under Grant N00014-91-J-1004 and by AFOSR under Grant F49620-98-1-0349.

* Corresponding author. Tel.: + 1-519-888-4567; fax: + 1-519-746-4791.

E-mail address: pfieguth@uwaterloo.ca (P.W. Fieguth).

0005-1098/01/$ - see front matter
PII: S0005-1098(00)00174-6

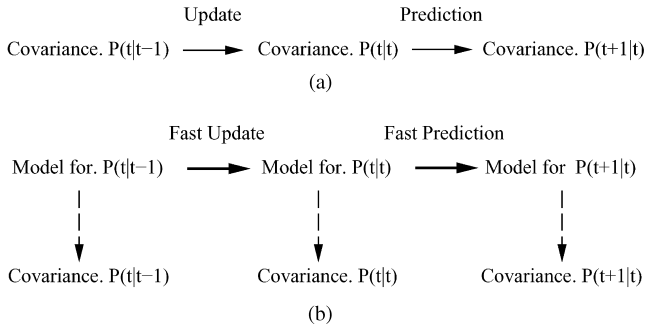


Fig. 1. Two possible sequences of steps for dynamic estimation: (a) the standard Kalman filter, in which covariance matrices are propagated; (b) proposed alternative, in which *models* are propagated.

errors, avoiding the storage or computation of large covariance matrices.

Consider time-recursive estimation for spatially distributed phenomena; this procedure can be viewed as an interleaved sequence of (i) temporal prediction and (ii) purely static spatial estimation problems. The standard Kalman filter approach, illustrated in Fig. 1(a), is to explicitly calculate the full covariance and the Kalman gain at each step. Each filter update step corresponds to solving a static estimation problem, namely that of estimating the errors in the one-step predicted estimates $\hat{x}(t|t-1)$ from the measurement innovations at time t. The exact Kalman filter propagation and solution of this problem corresponds to an *explicit* solution of each such static estimation problem by explicitly calculating full covariance and gain matrices, leading to associated complexity $\mathcal{O}(N^3)$.

An alternative recursive procedure (Chin, Karl, & Willsky, 1995) is one in which we propagate a statistical *model* for the one-step predicted estimation errors, as shown in Fig. 1(b). Such models *implicitly* specify the error statistics, although any desired element of the full error covariance can be computed (dashed lines in figure). The implicit nature of the representation leads to an *implicit* description of the optimal estimator; that is, our result is an algorithm rather than an explicit gain matrix, much as the Kalman filter is implicit and the Wiener filter explicit.

Clearly the major issue, then, is to find an implicit representation of the spatial error statistics that can be efficiently predicted and updated, improving on the $\mathcal{O}(N^3)$ complexity of the Kalman filter by orders of magnitude. In Chin et al. (1995) an approach was developed to use a Markov random field framework as an implicit representation for the spatial error models. Such models do indeed capture a rich class of spatial phenomena and in particular were demonstrated to lead to near-optimal estimation performance for problems in dynamic computer vision, however the actual solution of the spatial estimation problem for each measurement update using such a model is not nearly as efficient.

Instead, in this paper, we consider the use of an alternative implicit representation, namely the multiscale stochastic modeling and estimation methodology developed in Basseville et al. (1992) and Chou, Willsky and Benveniste (1994a). These multiscale models have been demonstrated to yield extremely fast solutions to purely spatial (i.e., temporally static estimation problems), including the modeling of $1/f$ processes (Daniel & Willsky, 1997a; Luettgen, Karl, Willsky, & Tenney, 1993), large distributed phenomena for remote sensing in oceanography (Fieguth et al., 1995; Fieguth, Menemenlis, Ho, Willsky, & Wunsch, 1998; Menemenlis, Fieguth, Wunsch, & Willsky, 1997) and hydrology (Daniel & Willsky, 1997b). For the class of multiscale models considered in this paper, given a multiscale model having a state dimension $d \ll N$, then the complexity to estimate a spatial process with N points is $\mathcal{O}(Nd^3)$, much less than $\mathcal{O}(N^3)$ for standard least-squares.

Taken together, the existence of the multiscale framework (a highly efficient static estimator) and the implicit modeling paradigm of Fig. 1 strongly motivate applying multiscale techniques to estimate *dynamic* or time-recursive systems; this is the fundamental contribution of this paper.

To apply this framework to a time-recursive problem requires finding a multiscale model for the estimation errors, and the derivation of an algorithm in order to propagate the multiscale model over time:

(1) Why should multiscale models be capable of modeling the estimation errors for distributed parameter systems? A rich literature already exists for the theory, stochastic realization, and parameter estimation of multiscale models for one-dimensional processes (Basseville et al., 1992; Chou et al., 1994a; Chou, Willsky, & Nikoukhah, 1994b; Daniel & Willsky, 1997b; Fieguth & Willsky, 1996; Irving, 1998; Luettgen & Willsky, 1995) and two-dimensional systems (Chin et al., 1995; Fieguth et al., 1995, 1998; Irving, Fieguth, & Willsky, 1997; Luettgen et al., 1993; Menemenlis et al., 1997).

We are interested in *approximating* the statistics of a given field; that is, we intentionally sacrifice a small amount of statistical fidelity in order to obtain multiresolution models that have small state dimension d. For a surprisingly rich class of purely spatial processes, low-dimensional multiresolution models have been constructed that yield near-optimal estimation performance (that is, with statistically insignificant discrepancies).

The basics of multiscale modeling are discussed in Sections 2 and 4.

(2) Past work on multiscale models considered static estimation problems, where the model is fixed a priori. In the time-recursive context the estimation error statistics, and consequently the associated

multiscale model, can change at each successive time if we are not in temporal steady-state. The particular problem of quickly propagating a multiscale model over time is unexplored and represents a new, significant contribution of this paper.

The principles of multiscale-model prediction are presented in Section 3. A new multiscale model, which leads to significant performance increases in the time-recursive context, is discussed in Sections 2 and 4, with examples and performance comparisons shown in Section 5.

2. Multiscale modeling and realization

In the multiscale estimation framework of Basseville et al. (1992), Chou et al. (1994a) and Irving (1998), random processes and random fields are modeled on tree structures. The nodes of these trees are organized into a sequence of scales, where the finest-level scale should be thought of as a discretization of the spatial domain of interest. A node s on the tree is connected to a unique parent node, $s\bar{\gamma}$, at the next coarser level, and to several child nodes $s\alpha_i$ $(i = 1, \ldots, q)$, at the next finer level. In general the number of children may vary from node to node. However, for our purposes focusing on the 1-D spatial domain, it is sufficient for us to restrict our attention to uniform $q = 2$ dyadic trees, depicted in Fig. 2.

The multiscale process is a collection of zero-mean random vectors $\mathbf{x}(s)$, indexed by nodes s on the tree and specified by a scale-to-scale relationship of the form

$$\mathbf{x}(s) = \mathbf{A}(s)\mathbf{x}(s\bar{\gamma}) + \mathbf{B}(s)\mathbf{w}(s), \tag{1}$$

where $\mathbf{w}(s)$ is a zero-mean unit-variance white noise process uncorrelated with $\mathbf{x}(0)$, the state at the root node of the tree. Measurements can be made at any node:

$$\mathbf{y}(s) = \mathbf{C}(s)\mathbf{x}(s) + \mathbf{v}(s), \tag{2}$$

where $\mathbf{v}(s)$ is white, zero-mean, and uncorrelated with the process $\mathbf{x}(s)$.

From (1), the whiteness of $\mathbf{w}(s)$ implies that the state $\mathbf{x}(s)$ conditionally decorrelates the $q + 1$ subtrees connected to node s. This Markovianity property of the multiscale tree admits efficient scale-recursive smoothing algorithms (Chou et al., 1994a,b), similar to the Rauch–Tung–Striebel smoothing algorithm (Rauch, Tung, & Striebel, 1965). The algorithm, summarized in the Appendix, is exact and has a computational complexity of $\mathcal{O}(k^3N)$, where k is the state dimension of $\mathbf{x}(s)$ and N is the number of nodes at the finest scale, in order to compute the estimates *and* error covariances at *all* nodes of the tree, compared to $\mathcal{O}(N^3)$ for the standard LLSE formalism.

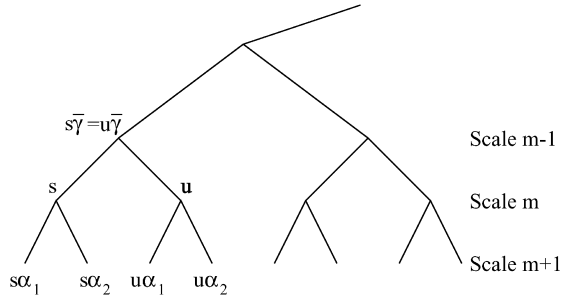


Fig. 2. A portion of a dyadic multiscale tree.

It is important to realize that the multiscale model (1), together with the covariance $\mathbf{P}(0)$ of state $\mathbf{x}(0)$ at the root of the tree, provides an *implicit* specification of the full covariance of the multiscale process. The explicit covariance between any two nodes $\mathbf{x}(s_1)$ and $\mathbf{x}(s_2)$ can be easily calculated as

$$\begin{aligned}\mathbf{P}(s_1, s_2) &\triangleq E[\mathbf{x}(s_1)\mathbf{x}^{\mathrm{T}}(s_2)] \\ &= \mathbf{\Phi}(s_1, s_1 \wedge s_2)\mathbf{P}(s_1 \wedge s_2)\mathbf{\Phi}^{\mathrm{T}}(s_2, s_1 \wedge s_2),\end{aligned} \tag{3}$$

where $s_1 \wedge s_2$ is the first common ancestor of s_1 and s_2, $\mathbf{P}(s_1 \wedge s_2)$ is the covariance of $\mathbf{x}(s_1 \wedge s_2)$, and $\mathbf{\Phi}(s,\sigma)$ is the state transition matrix from any node σ to direct descendent s. For example, referring to Fig. 2,

$$\mathbf{P}(s\alpha_i, u) \triangleq E[\mathbf{x}(s\alpha_i)\mathbf{x}^{\mathrm{T}}(u)] = \mathbf{A}(s\alpha_i)\mathbf{A}(s)\mathbf{P}(s\bar{\gamma})\mathbf{A}^{\mathrm{T}}(u). \tag{4}$$

Furthermore, the covariances of $\mathbf{x}(s)$ at each individual node can be recursively computed from a tree-recursive Lyapunov equation

$$\mathbf{P}(s) = \mathbf{A}(s)\mathbf{P}(s\bar{\gamma})\mathbf{A}^{\mathrm{T}}(s) + \mathbf{B}(s)\mathbf{B}^{\mathrm{T}}(s). \tag{5}$$

Thus the calculation of $\mathbf{P}(s)$ and any individual $\mathbf{P}(s_1, s_2)$ is computationally simple (at most $\mathcal{O}(N)$ for all of the $\mathbf{P}(s)$), whereas clearly the calculation of *all* of the cross-covariances $\mathbf{P}(s_1, s_2)$ is prohibitively complex.

The utility of the efficient estimation algorithm in the Appendix for the multiscale model (1) and (2) depends, of course, on the expressive power of models of this form. There exists a body of research (Luettgen et al., 1993; Daniel & Willsky, 1997b; Irving et al., 1997; Irving, 1998) for the stochastic realization of multiscale models that exactly or approximately match the second-order statistics of a given process.

Let $\boldsymbol{\chi}$ denote the ideal finest-scale process we wish to realize, and let $\mathbf{x}$ denote the subset of the state variables of a tree model we choose to use to model $\boldsymbol{\chi}$. That is, we wish to specify a tree model such that the covariance, $\mathbf{P}_{\mathbf{x}}$ approximately equals the ideal covariance $\mathbf{P}_{\boldsymbol{\chi}}$. As discussed in previous work on multiscale modeling, the realization problem consists of two distinct but related steps: (i) the specification of the state variables $\mathbf{x}(s)$ at every node; and (ii) the specification of the matrices $\mathbf{A}(s), \mathbf{B}(s)$.

The key in defining the coarser scale states $\mathbf{x}(s)$ is in satisfying the tree Markovianity property; that is, to

decorrelate the disjoint sets of variables on the several subtrees that node s separates. Defining each state vector $\mathbf{x}(s)$ as a linear functional of the process $\mathbf{x}$,

$$\mathbf{x}(s) = \mathbf{L}(s)\mathbf{x}, \tag{6}$$

a general method for constructing $\mathbf{L}(s)$ to achieve the required exact or approximate decorrelation is described in Irving (1998) and represents a generalization of Akaike's canonical correlations algorithm (Akaike, 1975). However in some cases, including ours, this step can be bypassed completely. As developed in Luettgen et al. (1993), for the case of first-order Markov process an exact multiscale realization can be constructed as shown in Fig. 3(a). This so-called *endpoint model* takes advantage of the fact that conditioned on the values at the two endpoints of any time interval, the remaining uncertainty about the values of a Markov process inside that interval is independent of those outside the interval. Since descendants of each node s in the figure only involve process values within the interval associated with node s, we immediately have tree-Markovianity.

In general, a model which collects the entire modeled process at the finest scale leads to considerable redundancy. In the case of the endpoint model just described, this redundancy is quite apparent, as the endpoints comprising the state at any parent node are copied into some of the descendent nodes. This suggests an alternative, Fig. 3(b), in which we eliminate this redundant copying and instead include endpoints of successively smaller intervals, leading to a *non-redundant* endpoint model. In this case, the process being realized has its sample values distributed over all of the tree nodes rather than having them all collected at the finest scale. While the existence of the non-redundant model illustrated in this figure represents a modest contribution to multiscale realization theory, its importance for the subject of this paper is far greater as its use greatly simplifies time-recursive estimation, as will be seen in Section 4.3.

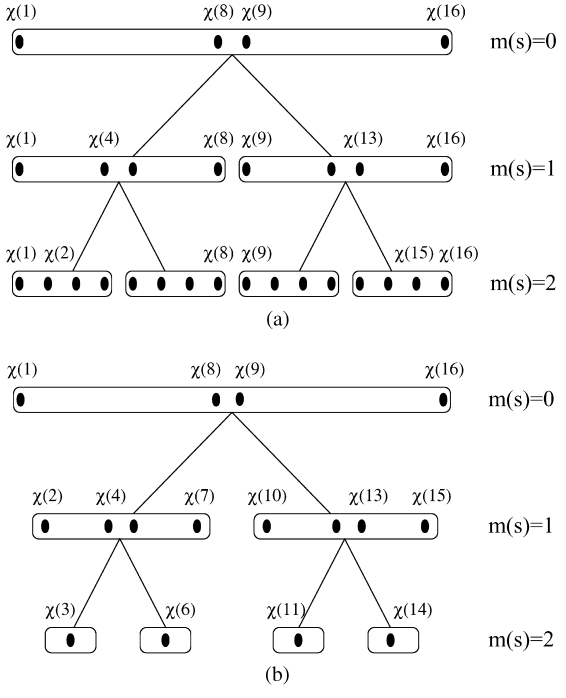


Fig. 3. Two possible 1-D Markov process realizations. (a) Endpoint model of Luettgen et al. (1993) for a first-order Markov process, here consisting of sixteen elements $\chi(1)\ldots\chi(16)$. Note that the entire set of samples is represented at the finest scale, but that some of these values are redundantly represented at nodes at higher levels in the tree. (b) Proposed nonredundant multiscale model for the same process, with the same root node, but with descendant states which are defined as endpoints of successively smaller intervals, eliminating redundancy. Note that each sample of χ appears only once in this tree model.

Once the state variables are specified, the second step of the realization problem is the construction of the dynamic matrices $\mathbf{A}(s)$,$\mathbf{B}(s)$. Since we can interpret (1) as the sum of an estimate $\mathbf{x}(s)$ based on its parent $\mathbf{x}(s\bar{\gamma})$ and the (orthogonal) error in this estimate, these matrices are completely determined by the joint statistics of $\mathbf{x}(s)$ and $\mathbf{x}(s\bar{\gamma})$:

$$\mathbf{A}(s) = \mathbf{P}(s, s\bar{\gamma})\mathbf{P}^{-1}(s\bar{\gamma}), \tag{7}$$

$$\mathbf{B}(s)\mathbf{B}^{\mathrm{T}}(s) = \mathbf{P}(s) - \mathbf{P}(s, s\bar{\gamma})\mathbf{P}^{-1}(s\bar{\gamma})\mathbf{P}^{\mathrm{T}}(s, s\bar{\gamma}). \tag{8}$$

Thus determining these matrices requires computing the joint parent–child statistics at each node. From (6), if we assume that the realization exactly matches the desired statistics, i.e., that $\mathbf{P}_{\chi} = \mathbf{P}_{\mathbf{x}}$, then

$$\mathbf{P}(s, s\bar{\gamma}) = \mathbf{L}(s)\mathbf{P}_{\chi}\mathbf{L}^{\mathrm{T}}(s\bar{\gamma}), \tag{9}$$

$$\mathbf{P}(s) = \mathbf{L}(s)\mathbf{P}_{\chi}\mathbf{L}^{\mathrm{T}}(s). \tag{10}$$

At first glance this seems prohibitive, as it implies that we need to explicitly calculate and store the entire covariance $\mathbf{P}_{\chi}$. However, from (7), (8) we really only need the second-order statistics of parent–child state pairs, which represent only low-dimensional projections of the full covariance. In particular, for the endpoint models in Fig. 3, the computation of $\mathbf{A}(s)$ and $\mathbf{B}(s)$ requires the specification of a relatively sparse subset of the elements of $\mathbf{P}_{\chi}$ of cardinality $\mathcal{O}(N)$ rather than $\mathcal{O}(N^2)$. It is this fact that leads us to an efficient solution to time-recursive multiscale estimation of diffusion processes.

3. Multiscale dynamic estimation

3.1. General approach

Consider a discrete-time system, whose temporal dynamics are governed by

$$\mathbf{z}(t+1) = \mathbf{A}_{\mathrm{d}}\mathbf{z}(t) + \mathbf{w}_{\mathrm{d}}(t), \tag{11}$$

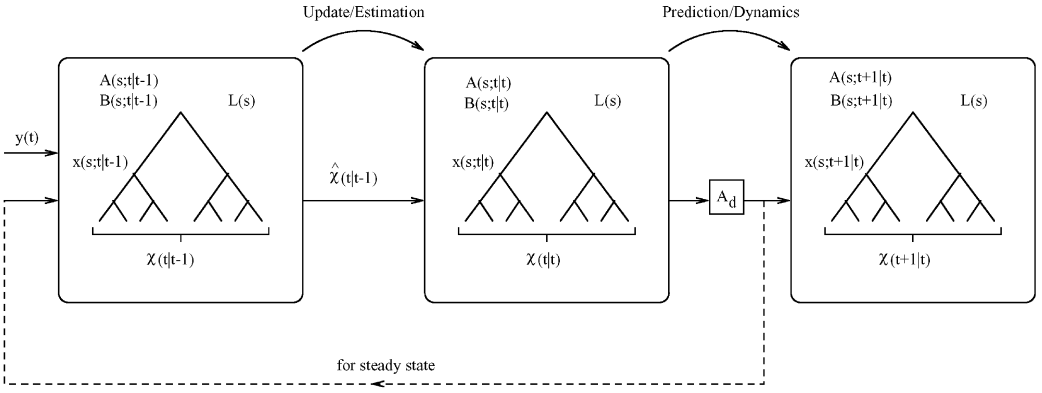


Fig. 4. A schematic of the proposed multiscale iterative method for dynamic estimation, modeled on the Kalman filter.

where $\mathbf{w}_d(t)$ is the zero mean process noise with diagonal covariance $\mathbf{Q}_d$. The measurements are

$$\mathbf{y}_d(t) = \mathbf{C}_d(t)\mathbf{z}(t) + \mathbf{v}_d(t), \tag{12}$$

where $\mathbf{v}_d(t)$ is the measurement noise with zero mean and diagonal covariance $\mathbf{R}_d$. The temporal dynamics, process noise, and measurement noise are assumed to be stationary in time, so $\mathbf{A}_d$, $\mathbf{Q}_d$, and $\mathbf{R}_d$ are independent of t. For the applications we have in mind (see Section 4) $\mathbf{z}(t)$ would represent a spatially discretized distributed parameter process and (11) would represent the corresponding temporally discretized dynamics, so that $\mathbf{A}_d$ represents the discretization of a partial differential operator in space. In addition, $\mathbf{R}_d$ is diagonal and the components of $\mathbf{y}_d$ represent independent point measurements of the distributed process.

We are interested in modeling the estimation error, so we let

$$\boldsymbol{\chi}(t|\tau) = \mathbf{z}(t) - \hat{\mathbf{z}}(t|\tau), \tag{13}$$

where $\hat{\mathbf{z}}(t|\tau)$ denotes the estimate of $\mathbf{z}(t)$ based on measurements through time τ. The Kalman filter, as sketched in Fig. 1, consists of a prediction stage

$$\hat{\mathbf{z}}(t+1|t) = \mathbf{A}_d\hat{\mathbf{z}}(t|t), \tag{14}$$

and a measurement update stage

$$\hat{\mathbf{z}}(t|t) = \hat{\mathbf{z}}(t|t-1) + \hat{\boldsymbol{\chi}}(t|t-1). \tag{15}$$

In standard Kalman filtering the estimate $\hat{\boldsymbol{\chi}}(t|t-1)$ is calculated explicitly as

$$\hat{\boldsymbol{\chi}}(t|t-1) = \mathbf{P}_{\chi}(t|t-1)\mathbf{C}_d^{\mathrm{T}}(\mathbf{C}_d\mathbf{P}_{\chi}(t|t-1)\mathbf{C}_d^{\mathrm{T}} + \mathbf{R}_d)^{-1}\mathbf{y}(t). \tag{16}$$

However, for the problems of interest here the dimensionality of $\boldsymbol{\chi}(t|t-1)$ makes this explicit calculation either impossible or at best exceedingly complex.

The alternative approach that we propose in this paper is to *implicitly* calculate and propagate the statistics of the estimation error as a sequence of multiscale models, illustrated in Fig. 4. Specifically, suppose that we have a multiscale model $\mathbf{A}(s;\, t|t-1)$, $\mathbf{B}(s;\, t|t-1)$ for the prediction error, defining the states as

$$\mathbf{x}(s;\, t|t-1) = \mathbf{L}(s)\boldsymbol{\chi}(t|t-1). \tag{17}$$

The multiscale estimation formulation in the appendix yields the estimates $\hat{\mathbf{x}}(s;\, t|t-1)$ *and* a multiscale model $\mathbf{A}(s;\, t|t)$, $\mathbf{B}(s;\, t|t)$ for the updated estimation error $\boldsymbol{\chi}(t|t)$:

$$\mathbf{x}(s;\, t|t) = \mathbf{L}(s)\boldsymbol{\chi}(t|t), \tag{18}$$

where the parameters $\mathbf{A}(s;\, t|t)$ and $\mathbf{B}(s;\, t|t)$ are computed as part of the multiscale estimation process. That is, if we start with a multiscale model (17) for $\boldsymbol{\chi}(t|t-1)$ we directly obtain an analogous model (18) for $\boldsymbol{\chi}(t|t)$ without explicitly calculating $\mathbf{P}_{\chi}(t|t)$.

To complete one step of the recursion we need to compute a multiscale model for the next predicted errors

$$\boldsymbol{\chi}(t+1|t) = \mathbf{A}_d\boldsymbol{\chi}(t|t) + \mathbf{w}_d(t) \tag{19}$$

without explicitly calculating $\mathbf{P}_{\chi}(t+1|t)$. Finding the predicted multiscale model, not provided by the multiscale estimation formulation, is explored in the following subsection and is novel to this paper.

3.2. Multiscale prediction step

We assume that the linear functionals $\mathbf{L}(s)$ have been specified and do not vary over time, although in general one might expect these linear functionals to change depending on how the statistics of the one-step prediction errors vary over time. With the choice of the linear functionals $\mathbf{L}(s)$ made, we are left with the final key issue, namely determining and propagating the parameters of the multiscale model through the temporal dynamics (19).

We assume that we know the predicted model $\mathbf{A}(s;\, t|t-1)$, $\mathbf{B}(s;\, t|t-1)$, $\mathbf{P}(s;\, t|t-1)$ in Fig. 4. Using the estimation and error modeling algorithms (Appendix) we can compute the corresponding quantities $\mathbf{A}(s;\, t|t)$, $\mathbf{B}(s;\, t|t)$, and $\mathbf{P}(s;\, t|t)$ for the updated multiscale model. Finally, we need to calculate the corresponding quantities $\mathbf{A}(s;\, t+1|t)$, $\mathbf{B}(s;\, t+1|t)$, and $\mathbf{P}(s;\, t+1|t)$ for the predicted model for $\boldsymbol{\chi}(t+1|t)$ whose states are

$$\mathbf{x}(s;\, t+1|t) = \mathbf{L}(s)\boldsymbol{\chi}(t+1|t). \tag{20}$$

From (7)–(10), we see that $\mathbf{A}(s;\, t+1|t)$ and $\mathbf{B}(s;\, t+1|t)$ will be determined if we determine both the individual state covariances $\mathbf{P}(s;\, t+1|t)$ and the parent–child cross-covariances $\mathbf{P}(s, \hat{\gamma};\, t+1|t)$. To derive expressions for the individual elements of these covariances, we first substitute the temporal dynamics (19) into (20):

$$\mathbf{x}(s;\, t+1|t) = \mathbf{L}(s)\mathbf{A}_{\mathrm{d}}\boldsymbol{\chi}(t|t) + \mathbf{L}(s)\mathbf{w}_{\mathrm{d}}(t). \tag{21}$$

Unless $\mathbf{A}_{\mathrm{d}}$ and $\mathbf{L}(s)$ commute, the term $\mathbf{L}(s)\mathbf{A}_{\mathrm{d}}\boldsymbol{\chi}(t|t)$ mixes the linear functionals used to form the states $\mathbf{x}(s;\, t|t)$ in (18). If we let $\mathbf{l}_i^{\mathrm{T}}(s)$ denote the ith row (linear functional) of $\mathbf{L}(s)$ then

$$x_i(s;\, t+1|t) = \mathbf{l}_i^{\mathrm{T}}(s)\mathbf{A}_{\mathrm{d}}\boldsymbol{\chi}(t|t) + \mathbf{l}_i^{\mathrm{T}}(s)\mathbf{w}_{\mathrm{d}}(t). \tag{22}$$

The term $\mathbf{l}_i^{\mathrm{T}}(s)\mathbf{A}_{\mathrm{d}}$ represents some linear functional of $\boldsymbol{\chi}(t|t)$, but in general it will not correspond to any of the linear functionals which we already have in $\mathbf{L}(s)$. However, it is always possible to write it as a *linear combination* of existing functionals, since the collection of linear functionals at the finest scale already forms a basis. Therefore we can write

$$\mathbf{l}_i^{\mathrm{T}}(s)\mathbf{A}_{\mathrm{d}} = \sum_{(\sigma,j)\in S} h_{\sigma,j}^{s,i}\mathbf{l}_j^{\mathrm{T}}(\sigma), \tag{23}$$

allowing us to express $x_i(s;\, t+1|t)$, not in terms of $\boldsymbol{\chi}(t|t)$, but instead in terms of selected model states:

$$x_i(s, t+1|t) = \sum_{(\sigma,j)\in S} h_{\sigma,j}^{s,i}x_j(\sigma, t|t) + \mathbf{l}_i^{\mathrm{T}}(s)\mathbf{w}_{\mathrm{d}}(t). \tag{24}$$

From (24) we can compute the quantities $\mathbf{P}(s;\, t+1|t)$ and $\mathbf{P}(s, \bar{\gamma};\, t+1|t)$ by computing certain covariances

$$E[x_i(s;\, t+1|t)x_j(u;t+1|t)], \tag{25}$$

which itself is computed from the known covariances $\mathbf{P}(s;\, t|t)$ and (3). However, as we pointed out in Section 2, calculating all or even many of these cross-covariances is prohibitive, thus it is desirable to choose among the various solutions to (23) those in which $h_{\sigma,j}^{s,i}$ are extremely sparse and in fact are nonzero only for nodes σ that are near to node s.

Since the specific properties of the $h_{\sigma,j}^{s,i}$ are highly dependent upon the dynamics, the next section will study the problem in a specific dynamic context.

4. Applications to 1-D diffusions

In this section we apply the ideas of Section 3 in detail to estimating 1-D diffusion problems. We will develop solutions to the two major issues identified in Section 3: the choice of the linear functionals $\mathbf{L}(s)$, and the propagation of the multiresolution model. In Section 5, we will illustrate the performance of the resulting estimator.

4.1. Problem setup

The point of departure for this application is 1-D damped heat diffusion process on a rod or ring satisfying the following stochastic PDE:

$$\frac{\partial z(l,\tau)}{\partial \tau} = a\,\frac{\partial^2 z(l,\tau)}{\partial l^2} - bz(l,\tau) + cw(l,\tau), \tag{26}$$

where $z(l,\tau)$ is the temperature at location l and time τ, $w(l,\tau)$ is a white Gaussian noise with unit variance, and $l\in[0,L]$. Constant a is related to the heat conduction coefficient and the problem dimension; b controls the heat loss to the surrounding coolant, whose temperature is set to zero without loss of generality.

The number of free parameters in (26) can be reduced by normalizing the spatial dimension to unit length and the diffusion parameter to 1:

$$\frac{\partial z(l,\tau)}{\partial \tau} = \frac{\partial^2 z(l,\tau)}{\partial l^2} - \beta z(l,\tau) + \gamma w(l,\tau). \tag{27}$$

A number of finite-difference schemes can be applied to discretize this PDE to arrive at a system of difference equations in the form of (11)

$$\mathbf{z}(t+1) = \mathbf{A}_{\mathrm{d}}\mathbf{z}(t) + \mathbf{w}_{\mathrm{d}}(t), \tag{28}$$

where $\mathbf{z}(t)$ is the vector containing the temperatures at all spatial grid points at time step t, and $\mathbf{w}_{\mathrm{d}}(t)$ models the process noise, with covariance $\mathbf{Q}_{\mathrm{d}}$. For our purposes such a discretized model plays two related but distinct roles: the prediction of the estimates (14), and to provide the dynamic matrix (23)–(24) to predict the estimation error statistics. As we will see, for the latter case it is desirable to choose $\mathbf{A}_{\mathrm{d}}$ to be banded with relatively small bandwidth, which arises if we use an explicit finite-difference temporal discretization. We use a simple forward Euler scheme, in which case $\mathbf{A}_{\mathrm{d}}$ is tridiagonal and $\mathbf{Q}_{\mathrm{d}} = \sigma_w^2\mathbf{I}$. Of course, if we use such a scheme prediction of estimates (14), care must be taken to ensure that the spatial discretization Δl and the temporal step $\Delta\tau$ are small enough for numerical accuracy and convergence (Strikwerda, 1989).

Obviously, a better choice for propagating estimates would be an implicit discretization scheme, which would result in a dense matrix $\mathbf{A}_{\mathrm{d}}$. We can actually consider using different $\mathbf{A}_{\mathrm{d}}$ for the two cases: an implicit, more "exact", scheme for the prediction of the estimates, and

an explicit scheme for the error model, propagating the error statistics only approximately. Consequently, the way in which new measurements are incorporated in the update step will not be optimal, however the multiscale error model *already* introduces an approximation into the update step, and we will see that the net effect of all of these approximations is a surprisingly small loss in performance.

We assume point measurements that may be irregular in space, but stationary in time (except for the last example in Section 5):

$$\mathbf{y}_{\mathrm{d}}(t) = \mathbf{C}_{\mathrm{d}}\mathbf{z}(t) + \mathbf{v}_{\mathrm{d}}(t), \tag{29}$$

where $\mathbf{C}_{\mathrm{d}}$ is a selection matrix and the measurement noise $\mathbf{v}_{\mathrm{d}}$ is white, covariance $\mathbf{R}_{\mathrm{d}} = \sigma_v^2\mathbf{I}$, uncorrelated with $\mathbf{z}(t)$ or $\mathbf{w}_{\mathrm{d}}(t)$. For a given particular measurement configuration $\mathbf{C}_{\mathrm{d}}$ we are left with only two free parameters: β and σ_w^2/σ_v^2.

Of course the complete specification of the model (28) also implies the specification of a specific set of boundary conditions. For the purpose of describing our methodology in this section, we will assume circular boundary conditions, $z(0,\tau) = z(L,\tau)$, physically corresponding to a thin cooling ring immersed in a coolant. In this case, the steady-state process variances are constant as long as the process (26) has spatially constant parameters; that is, the diagonal of $\mathbf{P}_{\mathbf{z}}$ is $\sigma_{\mathrm{p}}^2 I$. We will use the more familiar notion of signal-to-noise ratio $\mathrm{SNR} = 10\log(\sigma_{\mathrm{p}}^2/\sigma_v^2)$ instead of σ_w^2/σ_v^2. We can also adjust σ_w^2 to normalize σ_{p}^2 to 1. The stipulation of other boundary conditions leaves the linear functionals and the multiscale prediction algorithm unchanged and affects only the resulting numerical values.

4.2. Linear functionals

We are seeking a set of linear functionals which allow us to develop an accurate multiscale model for the steady-state predicted estimation errors in the context of one-dimensional diffusion. We propose to model the one-step predicted estimation errors as Markov processes, motivated by experimental work (Chin et al., 1995), which demonstrated cases in which estimation errors could be well-modeled by Markov random fields, and by theoretical work (Coleman, 1995), which showed that continuous-time, continuous-space heat diffusion models are Markov in steady-state.

In the specific case of one-dimensional Markov processes we have already seen in Fig. 3 an exact multiscale model (Chou et al., 1994a; Luettgen et al., 1993), in which the coarser scale states are defined as so-called "endpoint" linear functionals, for which each state consists of the finest-scale process values taken at state endpoints. Moreover, it has been demonstrated that such a choice of state is effective for many other processes as well (Daniel & Willsky, 1997a). Consequently, we will investigate the use of such functionals for space–time estimation problems of the type examined in our paper.

We will test our choice of linear functionals in two ways. We begin by deriving the *best* choice of linear functionals and compare these to the chosen endpoint functionals. Next, in Section 5, we will compare our multiscale approach, based on endpoint linear functionals, with the exact Kalman filter.

For small-size systems it is computationally possible to *explicitly* solve the Riccati equation for the exact steady-state error covariance, computing the full covariance $\mathbf{P}_{\mathbf{z}}(t|t-1)$ for the process to be realized at the finest level of the tree. We can then use a method of canonical correlations (CCR) (Irving, 1998), applying singular-value decompositions to parts of covariance $\mathbf{P}_{\mathbf{z}}(t|t-1)$, to produce at each node s a set of linear functionals ordered by statistical significance, measuring the degree to which a functional decorrelates node s from the remainder of the tree. In this way we can find the most appropriate selection of linear functionals to construct a multiscale realization. The insights gained from this procedure, applied to small-size systems, may then be applied to larger systems, where neither the Riccati equation nor CCR are computationally feasible.

Except in very special cases, the Riccati error covariance is spatially nonstationary and non-Markov as well. Nevertheless, the multiresolution representation still represents an excellent choice. Fig. 3 illustrates the application of CCR to decorrelating the interval $\{1-16\}$ from $\{17-32\}$ for two different measurement locations. Because of the intrinsic nonstationarity introduced by the tree, the location of the measurement has some influence on the results of CCR. The four most significant linear functionals produced by CCR are shown for the "best" (location 8) and "worst" (location 16) measurement placements (see Fig. 5). The immediate conclusion is that the two most significant linear functionals are almost completely concentrated on the interval endpoints. The relative insignificance of the third and fourth linear functionals as a function of measurement location is depicted in Fig. 6.

An alternative approach is to use the multiscale model itself, based on endpoint functionals, assessed via a *fractional variance reduction* (FVR) criterion, comparing the steady-state process variance and the steady-state updated error variance as a measure of estimator performance:

$$\mathrm{FVR} = \frac{\mathrm{Var(s.s.\ process)} - \mathrm{Var(s.s.\ updated\ error)}}{\mathrm{Var(s.s.\ process)}}. \tag{30}$$

For instance, if the FVR for the optimal estimator is 0.99 and for a suboptimal estimator is 0.98, we would argue that the suboptimal estimator has done a very good job, although its error variance is twice as large as the optimum.

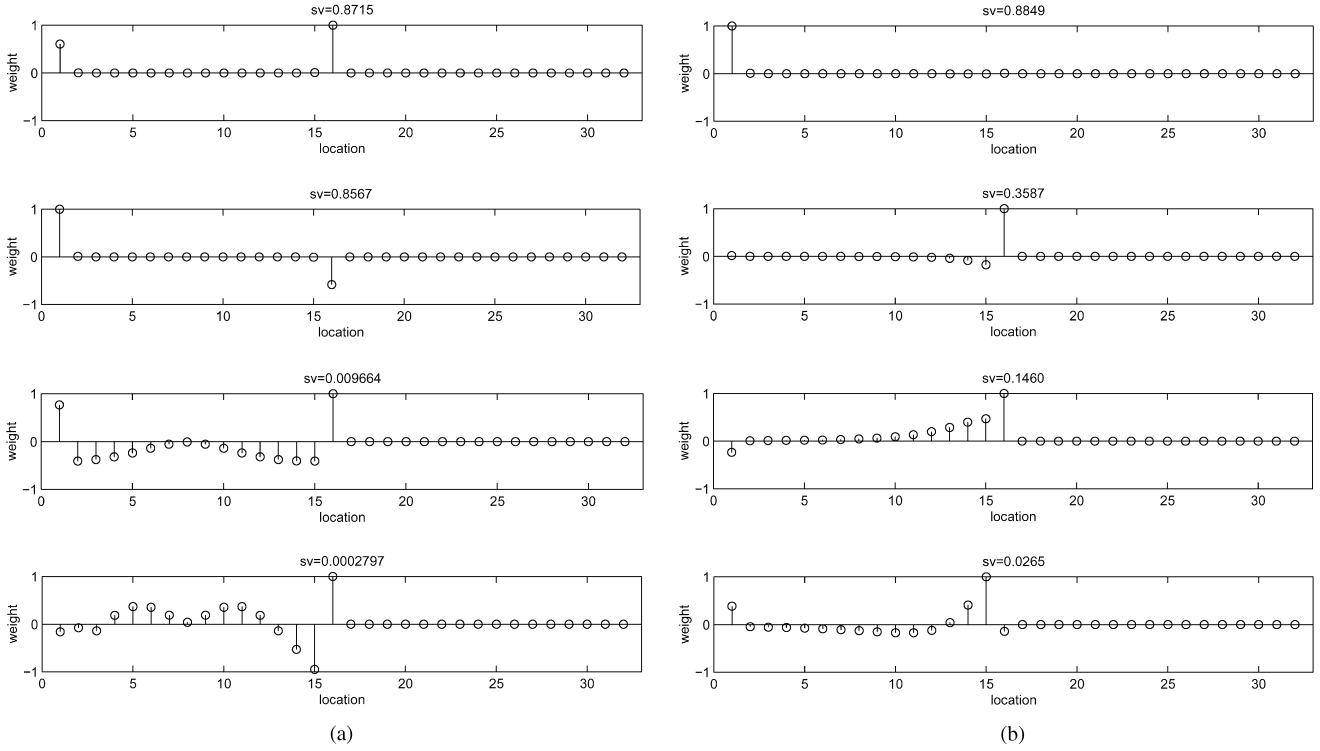


Fig. 5. The four most significant the linear functionals that decorrelate the steady-state predicted estimation errors at points 1–16 from those at points 17–32 of a 32-element diffusion process ($\beta = 10$, SNR $= 0$ dB, $\Delta\tau = 2 \times 10^{-5}$). The measurement is at pixel 8 in (a), and at pixel 16 in (b). The singular value is printed above each associated linear functional.

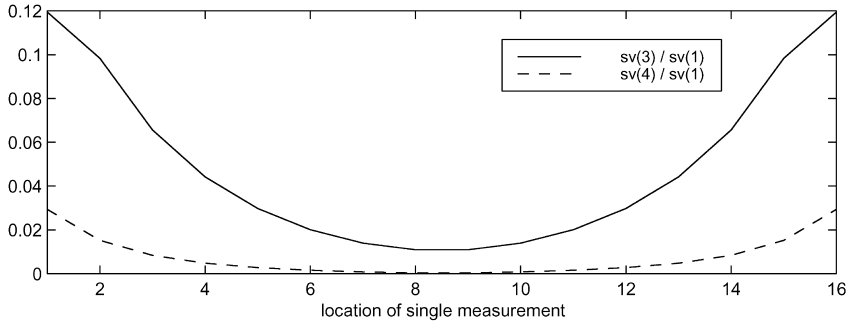


Fig. 6. The singular values of the third and fourth most significant linear functionals as a function of measurement location.

Fig. 7 depicts results for four different measurement locations. Fig. 7(a) shows the optimal and multiscale FVRs; at the resolution of this plot all of these curves are indistinguishable. Fig. 7(b) displays the percentage difference between each of the multiscale FVRs and the optimum; the differences are very small, peaking in the *worst-case* with an FVR of 0.596, whereas the optimal estimator has an FVR of 0.6.

4.3. Multiscale prediction step for 1-D diffusion

For the small examples considered thus far we could explicitly solve the Riccati equation, compute $\mathbf{P}_{\chi}(t+1|t)$, and determine the multiscale model for $\boldsymbol{\chi}(t+1|t)$. For large problems, however, we must directly infer the model for $\mathbf{P}_{\chi}(t+1|t)$ from the model for the updated errors $\mathbf{P}_{\chi}(t|t)$, which is computed by the multiscale estimation algorithm. The problem is that only the individual node covariances are explicitly calculated during multiscale estimation, whereas in general the mixing due to the dynamics $\mathbf{A}_{\mathrm{d}}$ requires that more distant correlations be calculated (as specified by the $h^{s,i}_{\sigma,j}$ in (23)–(24)).

However, for the diffusion processes of interest here we can construct the multiscale model for $\boldsymbol{\chi}(t+1|t)$ with very few additional calculations. In particular, as we argued in Section 4.2, the prediction error process

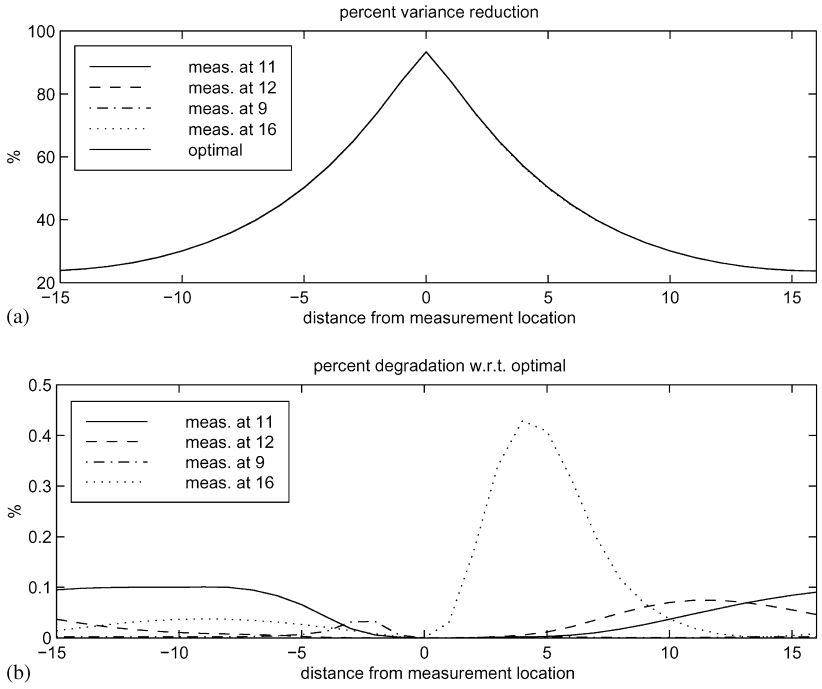


Fig. 7. (a) The percent variance reduction of the optimal estimator and of the multiscale estimator in steady state. (b) Percent degradation of the multiscale estimator with respect to the exact solution ($\beta = 10$, SNR $= 0$ dB, $\Delta\tau = 2 \times 10^{-5}$).

$\boldsymbol{\chi}(t+1|t)$ is well-approximated as a low-order Markov process in space, implying that we can represent it accurately using an endpoint multiscale model, most importantly a non-redundant endpoint model as in Fig. 3(b). Furthermore, for a diffusion process the matrix $\mathbf{A}_d$ is tridiagonal, implying that the dynamic evolution of the estimation error at any spatial location involves only the value of the error at that location and its nearest neighbors. As we can see from Fig. 3(b), the left and right spatial neighbors of *any* element of any node s can be found in s, the parent of s, or a child of s. Consequently, the statistics required to predict the multiscale model for $\boldsymbol{\chi}(t+1|t)$ are never more than three scales apart, regardless of the overall size of the problem. As a result, the complexity per tree node to construct this model is $\mathcal{O}(1)$, so that the total complexity is only $\mathcal{O}(N)$.

4.4. Iterative and recursive implementation

The complete algorithm we have just described can be used in one of the two ways. One, to obtain an approximate multiscale model for the steady-state prediction error process by running the algorithm iteratively off-line until convergence is achieved. Second, to use this algorithm to provide a multiscale error model dynamically at each step of the recursive estimation procedure for the initial, transient phase of estimation or for temporally non-stationary problems. The following paragraphs comment on issues of complexity, initialization, stopping criteria, and sources of inaccuracy.

The computational complexity per time-step of the algorithm is as follows. The end-point linear functionals have a state dimension ≤ 3 for any node on the tree, regardless of problem size. Therefore the total complexity of the update step is $\mathcal{O}(N)$. From the previous section, using the same linear functionals, the prediction step complexity is also $\mathcal{O}(N)$, therefore the total complexity of the dynamic multiscale estimator for discretized diffusion is only $\mathcal{O}(N)$ per time step!

The initialization of our algorithm takes the form of specifying a multiscale model for $\boldsymbol{\chi}(0)$, the prior estimation errors. Constructing such a model involves evaluating those elements of the prior covariance $\mathbf{P}_\chi(0)$ in order to derive the self-statistics of each tree node and the cross statistics between every node and its parent. While covariance extension and maximum-entropy methods (Dempster, 1972; Lev-Ari et al., 1989) can be used, often we can obtain these desired elements more easily using the FFT if the dynamics are space-invariant and assuming circular boundary conditions. This latter method is used for initialization in the examples in Section 5.

If we are iteratively calculating a multiscale model for the steady-state estimation errors, then the iteration stopping criterion is a critical issue. The convergence of the solution of the time-varying Riccati equation to steady-state is controlled by the slowest time constant of the error dynamics $\mathbf{A}_d(\mathbf{I} - \mathbf{K}(\infty)\mathbf{C}_d)$; choosing the number of iterations to be several times this time constant provides a conservative bound. For large problems this time constant will generally be unavailable, moreover

taking this conservative approach may lead to an excessive numbers of iterations. An alternative, adaptive, stopping criterion is to examine the diagonal elements of $\mathbf{P}_\chi(t|t)$ and stop when these suggest convergence, however in our experiments we restrict the number of iterations to be $\mathcal{O}(\log N)$ so that the total complexity is $\mathcal{O}(N \log N)$. Although this implies that the resulting multiscale estimator may not have converged, the results demonstrate that the performance of the resulting estimators is close to the optimal Kalman filter.

Given that the multiscale estimation algorithm of the Appendix is exact, the only sources of error lie in the realized multiscale model itself: the termination of iterations prior to convergence, the temporal and spatial discretization of the dynamics, and choosing end-point linear functionals as the basis for the multiscale model. Furthermore, because our model propagation assumes the updated statistics to be exact, it is possible that errors are accumulating over time. Although each of these sources of error can be reduced at the expense of additional computational complexity, the results of the following section will show that the algorithm performs nearly optimally at little statistical cost.

5. Examples and results

In this section we illustrate the application of our methodology to several examples of size $N = 64$. At this size exact calculations for the optimal estimator are still feasible, however the iterative multiscale algorithm described in this paper can in principle be applied more generally to much larger dynamic processes whose steady-state error process can be adequately modeled using end-point linear functionals and whose dynamics are local. This section will illustrate some extensions and departures from the basic diffusion problem described in the preceding section.

5.1. Cooling ring

We start with a cooling ring and a single measurement. The steady-state process variance has been normalized to 1. Fig. 8(a) shows the variance reduction plots for several values of SNR and heat loss parameter $\beta = 10$. As the SNR increases, so does the percent variance reduction. In all cases, the multiscale estimator is less than 0.2% poorer than the optimal estimator in steady state. The greatest degradation in performance occurs in regions furthest away from the measurement, where the error variances are large.

The single measurement case is, in a sense, the worse-case scenario, as the system is only weakly observable. In multiple measurement cases, the performance of our multiscale estimator compared to the optimal is generally better than that shown in Fig. 8 (Ho, 1998).

5.2. Pinned fin

In this example we replace the cyclic boundary condition by the more realistic condition for a cooling fin: one end of the fin is pinned to a heat source and the other end immersed in a coolant. The boundary condition at the heat source is Dirichlet: $z(0,t) = z_0$. At the free end, the heat flux is set to be equal to the heat loss, $\partial z(l,t)/\partial l = -\beta z(l,t)$. A second variation recognizes the fact that the heat loss parameter β may be spatially varying if the coolant is non-homogeneous (Aihara,

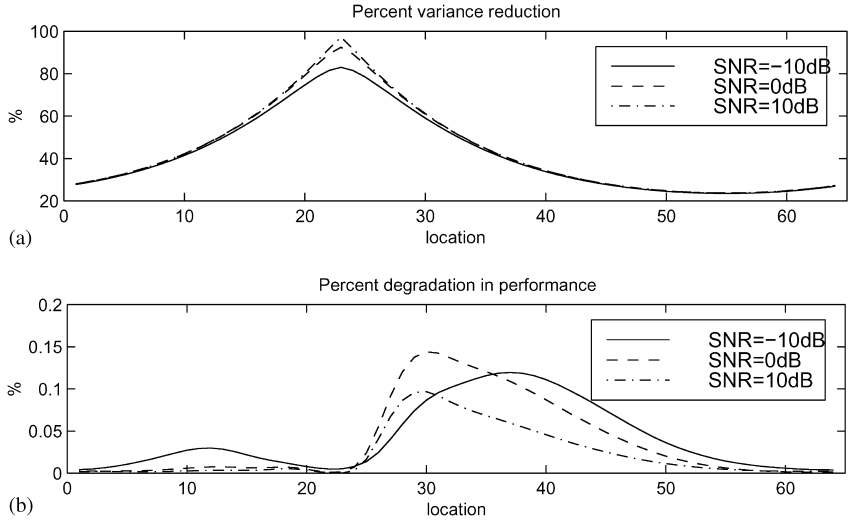


Fig. 8. (a) Percent variance reduction of the optimal estimator with one measurement at location 23 for $\beta = 10$ and SNR $= -10$, 0, and 10 dB. (b) Percent performance degradation with respect to optimal.

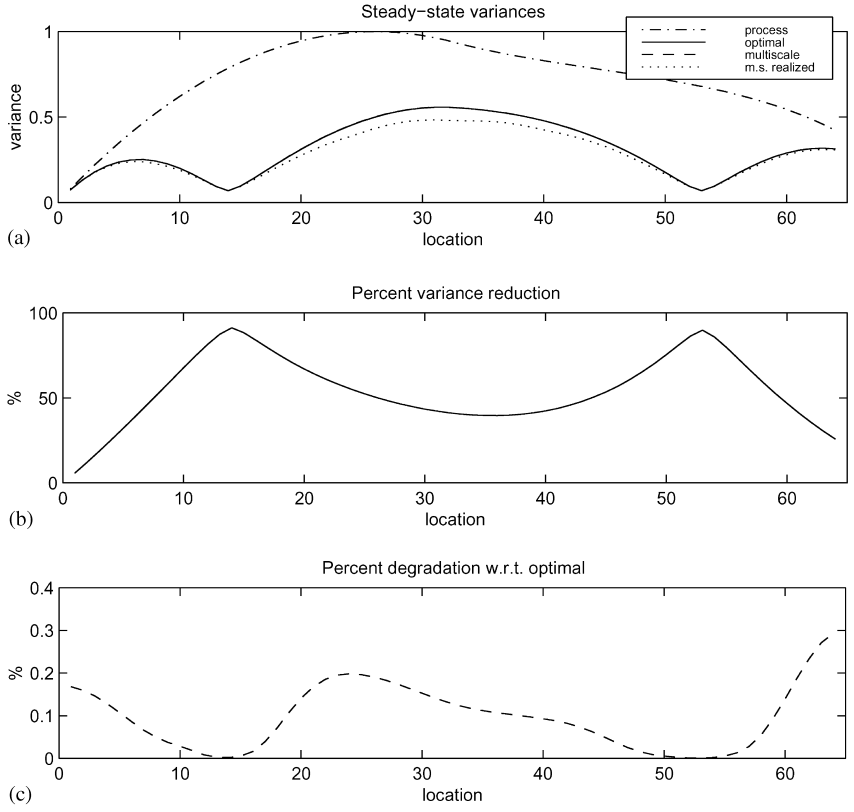


Fig. 9. Pinned fin with two measurements at locations 14 and 53. Heat loss parameter $\beta = 0$ at locations 1–32 and $\beta = 10$ at 33–64. (a) Steady-state process variances, steady-state estimation error variances, and the realized variances at the finest scale of the suboptimal multiscale estimator. (b) Percent variance reduction of the steady-state optimal estimators (SNR $= 0$ dB). (c) Percent performance degradation of the multiscale estimator with respect to optimal.

1997) (e.g., a partially insulated/partially cooled fin). The discretized dynamic equation (28) will then have a non-circulant $\mathbf{A}_d$ and an extra term $\mathbf{B}_d\mathbf{u}(t)$ to account for the boundary condition:

$$\mathbf{z}(t+1) = \mathbf{A}_d\mathbf{z}(t) + \mathbf{B}_d\mathbf{u}(t) + \mathbf{w}_d(t). \tag{31}$$

With non-circular boundary conditions or spatially-varying heat loss β the steady-state process becomes spatially non-stationary, requiring a modified definition of SNR; we will use the maximum pointwise SNR, $10\log(\max_i \sigma_p(i)^2/\sigma_v^2)$.

Fig. 9(a) shows the spatially nonstationary steady-state process variance and the steady-state error variances of the estimators for the case in which measurements are available at two spatial locations. The plots in panel (c) of the latter two are indistinguishable here: their difference is only a fraction of a percent, demonstrating the excellent performance of our method. Of course for truly large problems we would not have access to the optimal estimates nor the *actual* error variances of our suboptimal estimator. What we *do* have, however, is the multiscale model for $\boldsymbol{\chi}(t|t)$, the variances that this estimator *believes* it is achieving. This is illustrated as the dotted line in panel (a). Note that these variances are also quite accurate, although they slightly underestimate the actual error variance.

Fig. 10 shows the same pinned fin example of Fig. 9 except that a measurement is available only every 200 prediction steps, thus allowing substantial mixing to occur between measurements. Since *many* predictions steps must be taken for every update step, the effects of the approximations in our multiscale algorithm are much more pronounced, yet in the worse case it is still within 3% of the optimal estimator.

5.3. Advection–diffusion

Our multiscale methodology is also capable of modeling advection–diffusion processes, which have been employed in a wide variety of applications, especially in fluid dynamics, from pollution monitoring (Omatu et al., 1988) to tracer movements in oceanography (Wunsch, 1988, 1987).

The resulting dynamics

$$\frac{\partial z(l,\tau)}{\partial \tau} = \frac{\partial^2 z(l,\tau)}{\partial l^2} + \rho\frac{\partial z(l,\tau)}{\partial l} - \beta(l)z(l,\tau) + \gamma w(l,\tau) \tag{32}$$

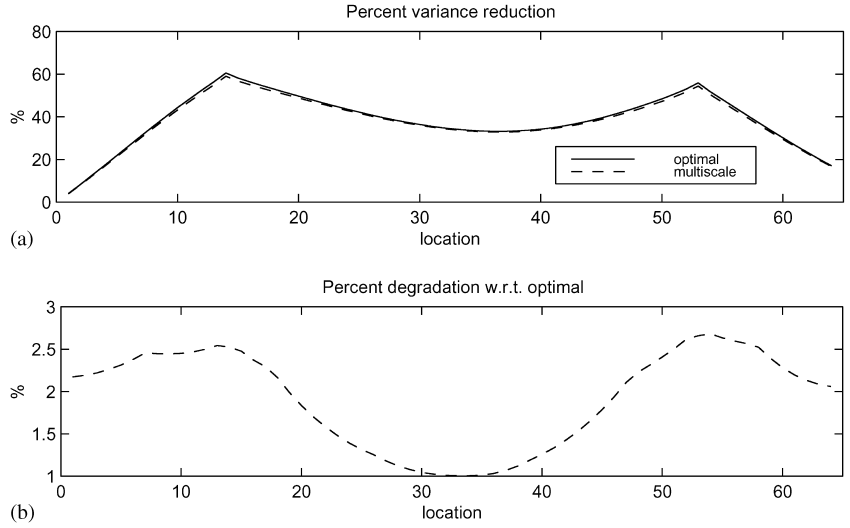


Fig. 10. As in Fig. 9, but with one measurement update every 200 prediction steps. (a) Percent variance reduction of the steady-state estimators. (b) Percent performance degradation of the multiscale estimator with respect to the optimal.

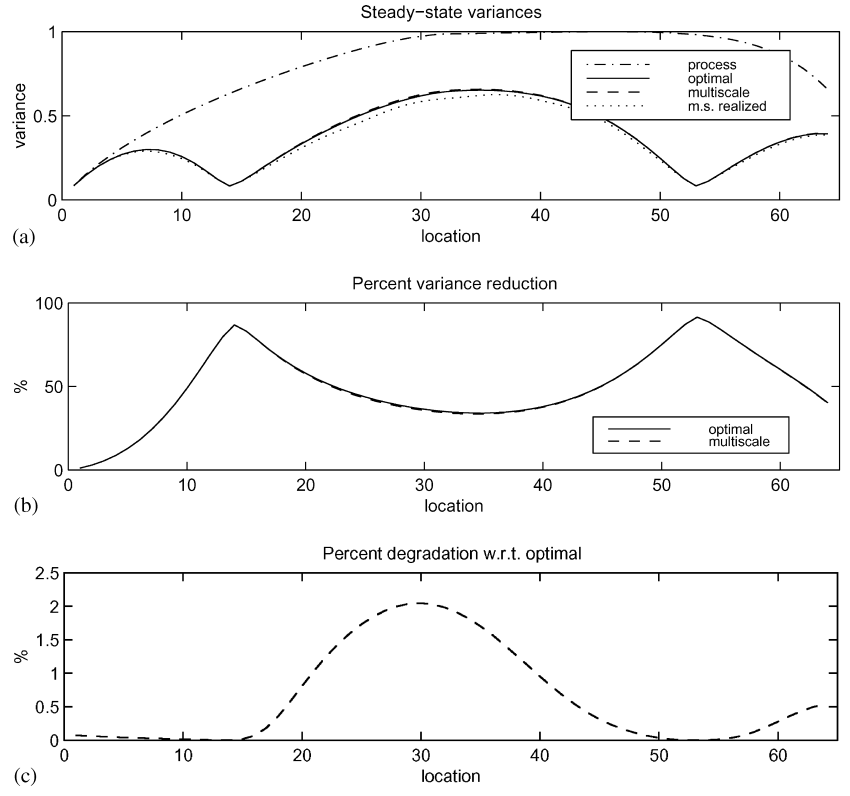


Fig. 11. Liquid flow ($\rho = -10$) from a reservoir (location 1) through a thin pipe, half insulated ($\beta = 0$ at locations 1–32) and half exposed ($\beta = 10$ at 33–64). Measurements at locations 14 and 53. SNR = 0 dB. (a) Steady-state process variances, steady-state estimation error variances, realized variances at the finest scale of the multiscale estimator. (b) Percent variance reduction of the steady-state estimators. (c) Percent performance degradation of the multiscale estimator with respect to the optimal.

model a thin pipe, in which a liquid flows towards positive l from a reservoir. Fig. 11 displays the results from one such estimation problem with measurements at two spatial locations. The multiscale approximate estimator tracks the performance of the optimal estimator closely. Also, the approximate error variances captured by the multiscale model (corresponding to the dotted line in panel (a)) provide a very good approximation to the actual error statistics.

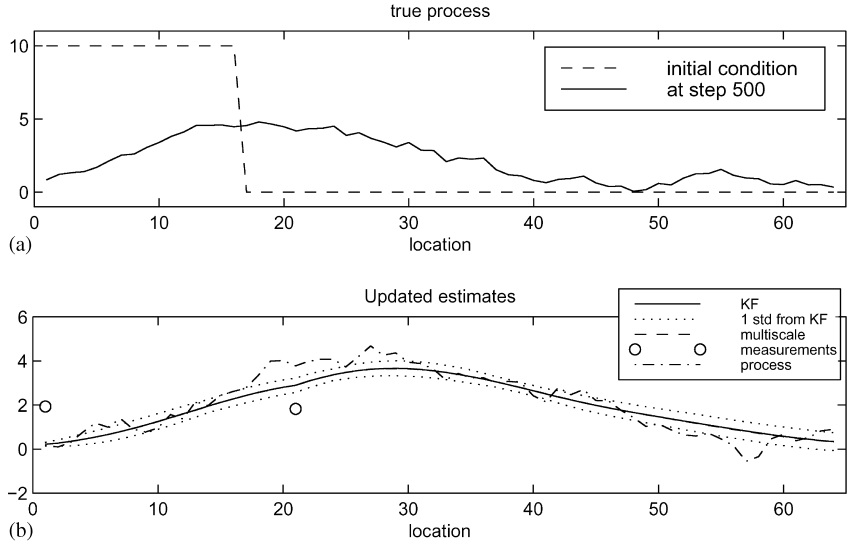


Fig. 12. Cooling pipe. One update step for every 100 prediction steps (SNR = 0 dB). (a) True process at step 1300 and initial values of the process. (b) Updated estimates at step 1300. The optimal and the multiscale suboptimal estimates are indistinguishable at this resolution. The locations and values of the two measurements at this update step are labeled with circles. Dotted curves show the range of one standard deviation from the optimal estimates.

5.4. Recursive implementation and temporally nonstationary performance

Fig. 12 illustrates the application of the recursive version of our algorithm to a temporally nonstationary situation, based on the advection–diffusion dynamics in (32), initializing the process with the nonequilibrium initial condition indicated by the solid line in Fig. 12(a). The measurements, taken once every 100 prediction steps, are nonstationary: at each measurement time the number of measurements is Poisson (mean 4) and the measurement locations are uniformly distributed.

Since there is no steady-state, we have depicted a snapshot of the process and estimation results at time step 1300 (i.e., after the 13th update). The dash-dot line in Fig. 12(a) indicates the actual process at time 1300, while Fig. 12(b) depicts the estimation results after the measurement update. The two measurements taken at this time are indicated by the small circles. As these figures illustrate, the estimates produced by the optimal Kalman filter and by our multiscale recursive estimator are virtually identical and the differences are statistically insignificant.

6. Conclusion

In this paper we have developed a new approach to suboptimal estimation for recursive estimation for distributed parameter space–time phenomena. The point of departure for our work are the basic equations of Kalman filtering, which can be prohibitively complex because of the growth in computational complexity with the dimension of the problem of interest. Indeed this is one of the most significant challenges faced in remote sensing data assimilation.

Our solution to this problem involves making use of the observation that each update step in recursive estimation can be viewed as a static estimation problem, in which the errors in the predicted estimates are estimated based on the latest measurement innovations. Rather than explicitly propagating the full error covariance for this prediction error field, we consider propagating a *model*. In particular, rather than using standard models such as Markov random fields, we have chosen to use a recently introduced class of multiscale models, which leads to extremely fast algorithms for estimation. The major challenge in applying this multiscale methodology is in developing a method for propagating multiscale error field models through the mixing introduced by the temporal dynamics of the process being estimated.

The estimation results obtained indicate that near-optimal performance can be achieved using this methodology. Indeed, we would argue that, compared to the intrinsic model uncertainty in many of the space–time processes of interest such as remote sensing, the differences in performance between our algorithm and the Kalman filtering solution are insignificant.

While we have illustrated our results here for 1-D spatial processes, much greater benefits can be expected in two- and three-dimensional problems. While the basic concept of how to develop this extension is described in this paper, important issues remain in order to make this extension a reality. In particular, the choice of multiscale

states in the representation of estimation error fields represents a first important problem that is currently under investigation. In addition, while diffusion and advection–diffusion problems such as those considered in this paper are of considerable practical interest in higher dimensions, it is also of considerable interest to understand how to adapt our methodology to dynamics that allow wave-like behavior. Obviously, for such models we would expect that the propagation of error models over time would need to account for the modes of wave propagation. Issues such as these as well as developing a deeper understanding of how to capture temporal mixing of scales within our multiresolution framework represent clearly defined directions to be pursued in order to fully realize the promise suggested by the results presented in this paper.

Appendix. Multiscale smoothing algorithm

The essential equations of the multiscale smoothing algorithm are listed here. More detailed development of these equations can be found in Chou et al. (1994a) and Luettgen and Willsky (1995).

Suppose that we are given the multiscale process and measurement equations:

$$\mathbf{x}(s) = \mathbf{A}(s)\mathbf{x}(s\bar{\gamma}) + \mathbf{B}(s)\mathbf{w}(s), \tag{A.1}$$

$$\mathbf{y}(s) = \mathbf{C}(s)\mathbf{x}(s) + \mathbf{v}(s), \tag{A.2}$$

where $\mathbf{w}(s)$ is a zero-mean unit-variance white-noise process and $\mathbf{v}(s)$ is a zero-mean white noise process with covariance $\mathbf{R}(s)$. We are also given the statistics of the states at the root node: zero mean with covariance $\mathbf{P}(0)$. First, the prior covariances of all states at individual nodes on the tree are computed via a Lyapunov equation

$$\mathbf{P}(s) = \mathbf{A}(s)\mathbf{P}(s\bar{\gamma})\mathbf{A}^{\mathrm{T}}(s) + \mathbf{B}(s)\mathbf{B}^{\mathrm{T}}(s). \tag{A.3}$$

The core of the multiscale algorithm consists of an upward estimation sweep and a downward smoothing sweep, but first let us define a few quantities:

$$\mathbf{Y}_s = \{\mathbf{y}(\sigma) | \sigma \text{ is a descendant of } s\}, \tag{A.4}$$

$$\hat{\mathbf{x}}(\sigma|s) = E[\mathbf{x}(\sigma)|\sigma \in \mathbf{Y}_s \cup \mathbf{y}(s)], \tag{A.5}$$

$$\hat{\mathbf{x}}(\sigma|s+) = E[\mathbf{x}(\sigma)|\sigma \in \mathbf{Y}_s], \tag{A.6}$$

$$\tilde{\mathbf{P}}(\sigma|s) = \mathrm{Cov}[\mathbf{x}(\sigma) - \hat{\mathbf{x}}(\sigma|s)], \tag{A.7}$$

$$\tilde{\mathbf{P}}(\sigma|s+) = \mathrm{Cov}[\mathbf{x}(\sigma) - \hat{\mathbf{x}}(\sigma|s+)]. \tag{A.8}$$

The upward sweep initializes at the finest level from the prior covariances:

$$\hat{\mathbf{x}}(s|s+) = 0, \tag{A.9}$$

$$\mathbf{P}(s|s+) = \mathbf{P}(s). \tag{A.10}$$

It requires the following upward model, corresponding to the the downward model in (A.1):

$$\mathbf{x}(s\bar{\gamma}) = \mathbf{F}(s)\mathbf{x}(s) + \bar{\mathbf{w}}(s), \tag{A.11}$$

$$\mathbf{y}(s) = \mathbf{C}(s)\mathbf{x}(s) + \mathbf{v}(s), \tag{A.12}$$

where

$$\mathbf{F}(s) = \mathbf{P}(s\bar{\gamma})\mathbf{A}^{\mathrm{T}}(s)\mathbf{P}(s)^{-1}, \tag{A.13}$$

$$E[\bar{\mathbf{w}}(s)\bar{\mathbf{w}}(s)^{\mathrm{T}}] = \mathbf{P}(s\bar{\gamma}) - \mathbf{P}(s\bar{\gamma})\mathbf{A}^{\mathrm{T}}(s)\mathbf{P}(s)^{-1}\mathbf{A}(s)\mathbf{P}(s\bar{\gamma}) = \mathbf{Q}(s). \tag{A.14}$$

The upward sweep computes the best estimate of the states at a node given all measurement below that node. It consists of three steps at each scale:

(a) *Update step*:

$$\hat{\mathbf{x}}(s|s) = \hat{\mathbf{x}}(s|s+) + \mathbf{K}(s)[\mathbf{y}(s) - \mathbf{C}(s)\hat{\mathbf{x}}(s|s+)], \tag{A.15}$$

$$\mathbf{P}(s|s) = [\mathbf{I} - \mathbf{K}(s)\mathbf{C}(s)]\mathbf{P}(s|s+), \tag{A.16}$$

$$\mathbf{K}(s) = \mathbf{P}(s|s+)\mathbf{C}^{\mathrm{T}}(s)[\mathbf{C}(s)\mathbf{P}(s|s+)\mathbf{C}^{\mathrm{T}}(s) + \mathbf{R}(s)]^{-1}. \tag{A.17}$$

(b) *Prediction step*:

$$\hat{\mathbf{x}}(s|s\alpha_i) = \mathbf{F}(s\alpha_i)\hat{\mathbf{x}}(s\alpha_i|s\alpha_i), \tag{A.18}$$

$$\mathbf{P}(s|s\alpha_i) = \mathbf{F}(s\alpha_i)\mathbf{P}(s\alpha_i|s\alpha_i)\mathbf{F}^{\mathrm{T}}(s\alpha_i) + \mathbf{Q}(s\alpha_i). \tag{A.19}$$

(c) *Merge step*:

$$\hat{\mathbf{x}}(s|s+) = \mathbf{P}(s|s+)\sum_{i=1}^{q} \mathbf{P}^{-1}(s|s\alpha_i)\hat{\mathbf{x}}(s|s\alpha_i), \tag{A.20}$$

$$\mathbf{P}(s|s+) = \left[(1-q)\mathbf{P}(s)^{-1} + \sum_{i=1}^{q} \mathbf{P}^{-1}(s|s\alpha_i)\right]^{-1}. \tag{A.21}$$

The downward sweep computes the best estimate of the states at a node given all available measurements:

$$\hat{\mathbf{x}}(s|0) = \hat{\mathbf{x}}(s|s) + \mathbf{J}(s)[\hat{\mathbf{x}}(s\bar{\gamma}|0) - \hat{\mathbf{x}}(s\bar{\gamma}|s)], \tag{A.22}$$

$$\mathbf{P}(s|0) = \mathbf{P}(s|s) + \mathbf{J}(s)[\mathbf{P}(s\bar{\gamma}|0) - \mathbf{P}(s\bar{\gamma}|s)]\mathbf{J}^{\mathrm{T}}(s), \tag{A.23}$$

$$\mathbf{J}(s) = \mathbf{P}(s|s)\mathbf{F}^{\mathrm{T}}(s)\mathbf{P}^{-1}(s\bar{\gamma}|s). \tag{A.24}$$

The smoothing error can be modeled as

$$\tilde{\mathbf{x}}(s|0) = \mathbf{J}(s)\tilde{\mathbf{x}}(s\bar{\gamma}|0) + \check{\mathbf{w}}(s), \tag{A.25}$$

where $\tilde{\mathbf{x}}(s|0) = \mathbf{x}(s) - \hat{\mathbf{x}}(s|0)$, and

$$E[\check{\mathbf{w}}(s)\check{\mathbf{w}}(s)^{\mathrm{T}}] = \mathbf{P}(s|s) - \mathbf{P}(s|s)\mathbf{F}^{\mathrm{T}}(s)\mathbf{P}^{-1}(s\bar{\gamma}|s)\mathbf{F}(s)\mathbf{P}(s|s). \tag{A.26}$$

Note that the state covariances at individual nodes of the smoothing error model have already been computed in (A.24) and (A.26).

References

Aihara, S. I. (1997). On adaptive boundary control for stochastic parabolic systems. *IEEE Transactions on Automatic Control*, *42*(3), 350–363.

Akaike, H. (1975). Markovian representation of stochastic processes by canonical variables. *SIAM Journal of Control*, *13*(1), 162–173.

Banks, H. T., & Kunisch, K. (1989). *Estimation techniques for distributed parameter systems*. Boston: Birkhäuser.

Basseville, M., Benveniste, A., Chou, K., Golden, S., Nikoukhah, R., & Willsky, A. (1992). Modeling and estimation of multiresolution stochastic processes. *IEEE Transactions on Information Theory*, *38*(2), 766–784.

Briggs, W. L. (1987). *A multigrid tutorial*. Philadelphia: Society for Industrial and Applied Mathematics.

Chin, T. M., Karl, W. C., & Willsky, A. S. (1995). A distributed and iterative method for square root filtering in space-time estimation. *Automatica*, *31*(1), 67–82.

Chou, K., Willsky, A., & Benveniste, A. (1994a). Multiscale recursive estimation, data fusion, and regularization. *IEEE Transactions on Automatic Control*, *39*(3), 464–478.

Chou, K., Willsky, A., & Nikoukhah, R. (1994b). Multiscale systems, Kalman filters, and Riccati equations. *IEEE Transactions on Automatic Control*, *39*(3), 479–492.

Coleman, J. M. (1995). *Gaussian spacetime models: Markov field properties*. Ph.D. thesis, University of California at Davis, David, CA.

Daniel, M., & Willsky, A. (1997a). Modeling and estimation of fractional Brownian motion using multiresolution stochastic processes. In Vehel, J. L., Lutton, E., and Tricot, C. (Eds.), *Fractals in engineering* (pp. 124–137). Berlin: Springer.

Daniel, M., & Willsky, A. (1997b). A multiresolution methodology for signal-level fusion and data assimilation with applications to remote sensing. *Proceedings of the IEEE*, *85*(1), 164–183.

Dempster, A. P. (1972). Covariance selection. *Biometrics*, *28*, 157–175.

Fieguth, P., Karl, W., Willsky, A., & Wunsch, C. (1995). Multiresolution optimal interpolation and statistical analysis of TOPEX/POSEIDON satellite altimetry. *IEEE Transactions on Geoscience and Remote Sensing*, *33*(2), 280–292.

Fieguth, P., Menemenlis, D., Ho, T., Willsky, A., & Wunsch, C. (1998). Mapping Mediterranean altimeter data with a multiresolution optimal interpolation algorithm. *Journal of Atmospheric and Oceanic Technology*, *15*, 535–546.

Fieguth, P., & Willsky, A. (1996). Fractal estimation using models on multiscale trees. *IEEE Transactions on Signal Processing*, *44*(5), 1297–1300.

Ho, T. T. (1998). *Multiscale modeling and estimation of large-scale dynamic systems*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Irving, W. W. (1998). A canonical correlations approach to multiscale stochastic realization. *IEEE Transactions on Automatic Control*, in press.

Irving, W. W., Fieguth, P. W., & Willsky, A. S. (1997). An overlapping tree approach to multiscale stochastic modeling and estimation. *IEEE Transactions on Image Processing*, *6*(11), 1517–1529.

Lev-Ari, H., Parker, S. R., & Kailath, T. (1989). Multidimensional maximum-entropy covariance extension. *IEEE Transactions on Information Theory*, *35*(3), 497–508.

Luettgen, M., Karl, W., Willsky, A., & Tenney, R. (1993). Multiscale representations of Markov random fields. *IEEE Transactions on Signal Processing*, *41*(12), 3377–3395.

Luettgen, M., & Willsky, A. (1995). Multiscale smoothing error models. *IEEE Transactions on Automatic Control*, *40*(1), 173–175.

Menemenlis, D., Fieguth, P., Wunsch, C., & Willsky, A. (1997). Adaptation of a fast optimal interpolation algorithm to the mapping of oceanographic data. *Journal of Geophysical Research*, *102*(C5), 10 573–10 584.

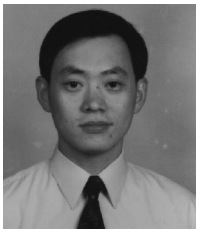
Omatu, S., Seifeld, J. H., Soeda, T., & Sawaragi, Y. (1988). Estimation of nitrogen dioxide concentration in the vicinity of a roadway by optimal filtering theory. *Automatica*, *24*(1), 19–29.

Rauch, H. E., Tung, F., & Striebel, C. T. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, *3*(8), 1445–1450.

Strikwerda, J. C. (1989). *Finite difference schemes and partial differential equations.*. New York: Chapman & Hall.

Wunsch, C. (1987). Using transient tracers: the regularization problem. *Tellus*, *39B*, 477–492.

Wunsch, C. (1988). Transient tracers as a problem in control theory. *Journal of Geophysical Research*, *93*(C7), 8099–8110.



Terrence T. Ho (S'88) received the B.E. degree from the Cooper Union for the Advancement of Science and Art, New York, in 1992, and the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, in 1998, both in electrical engineering. His Ph.D. research focused on multiscale estimation of large-scale time-varying problems. He was a recipient of the National Science Foundation Graduate Fellowship. He is a member of Tau Beta Pi and Eta Kappa Nu. He is currently with the Information Technology Group at Booz Allen & Hamilton in New York.

Paul W. Fieguth (S'87–M'96) received the B.A.Sc. degree from the University of Waterloo, Ontario, Canada, in 1991 and the Ph.D. degree from the Massachusetts Institute of Technology, Cambridge, in 1995, both degrees in electrical engineering. He joined the faculty at the University of Waterloo in 1996, where he is currently Assistant Professor in Systems Design Engineering. Prior to this appointment he held postdoctoral positions in the Department of Computer Science at the University of Toronto, and in the Laboratory for Information and Decision Systems at MIT. His research interests include statistical signal and image processing, hierarchical algorithms, data fusion, and interdisciplinary applications of such methods, particularly to remote sensing.

Alan S. Willsky received both the S.B. degree and the Ph.D. degree from the Massachusetts Institute of Technology in 1969 and 1973, respectively. He joined the M.I.T. faculty in 1973 and his present position is as the Edwin S. Webster Professor of Electrical Engineering. From 1974 to 1981 Dr. Willsky served as Assistant Director of the M.I.T. Laboratory for Information and Decision Systems. He is also a founder and member of the board of directors of Alphatech, Inc. and is also a member of the US Air Force Scientific Advisory Board. In 1975 he received the Donald P. Eckman Award from the American Automatic Control Council. Dr. Willsky has held visiting positions at Imperial College, London, L'Université de Paris-Sud, and the Institut de Recherche en Informatique et Systémes Aléatoires in Rennes, France. He was program chairman for the 17th IEEE Conference on Decision and Control, has been an associate editor of several journals including the *IEEE Transactions on Automatic Control*, has served as a member of the Board of Governors and Vice President for Technical Affairs of the IEEE Control Systems Society, was program chairman for the 1981 Bilateral Seminar on Control Systems held in the People's Republic of China, was special guest editor of the 1992 special issue of the *IEEE Transactions on Information Theory* on wavelet transforms and multiresolution signal analysis, and served as co-chair of the 1998 ONR Working Group on the Role of Probability and Statistics in Command and Control. Also in 1988 he was made a Distinguished Member of the IEEE

Control Systems Society. Dr. Willsky has given several plenary lectures at major scientific meetings including the 20th IEEE Conference on Decision and Control, the 1991 IEEE International Conference on Systems Engineering, the SIAM Conference on Applied Linear Algebra, 1991, the 1992 Inaugural Workshop for the National Centre for Robust and Adaptive Systems, Canberra, Australia, the 1993 IEEE Symposium on Image and Multidimensional Signal Processing, Cannes, France; and the 1997 SPIE Wavelets and Applications Symposium, San Diego.

Dr. Willsky is the author of the research monograph *Digital Signal Processing and Control and Estimation Theory* and is co-author of the undergraduate text *Signals and Systems*. He was awarded the 1979 Alfred Noble Prize by the ASCE and the 1980 Browder J. Thompson Memorial Prize Award by the IEEE for a paper excerpted from his monograph. Dr. Willsky's present research interests are in problems involving multidimensional and multiresolution estimation and imaging, statistical image and signal processing, data fusion and estimation of complex systems, image reconstruction, and computer vision.