Multiply-Rooted Multiscale Models for Large-Scale Estimation

Paul W. Fieguth (Member)

Abstract

Divide-and-conquer or multiscale techniques have become popular for solving large statistical estimation problems. The methods rely on defining a state which conditionally decorrelates the large problem into multiple subproblems, each more straightforward than the original. However this step cannot be carried out for asymptotically large problems since the dimension of the state grows without bound, leading to problems of computational complexity and numerical stability. In this paper we propose a new approach to hierarchical estimation in which the conditional decorrelation of arbitrarily-large regions is avoided, and the problem is instead addressed piece-by-piece. The approach possesses promising attributes: it is *not* a local method — the estimate at every point is based on *all* measurements; it is numerically stable for problems of arbitrary size; and the approach retains the benefits of the multiscale framework on which it is based: a broad class of statistical models, a stochastic realization theory, an algorithm to calculate statistical likelihoods, and the ability to fuse local and non-local measurements.

1 Introduction

The statistical estimation of large, global scale, two-dimensional remote sensing problems and even modestly-sized three-dimensional problems presents tremendous and pertinent challenges: heightened environmental awareness and concerns have led to an explosion in the quantity of remotely-sensed data, most of which contain irregular gaps and are governed by nonstationary underlying fields[24, 27]. That is, we are interested in extremely large estimation problems having nonstationary prior models.

Several approaches we dismiss out of hand: Brute-force, relying on full matrix inversion, is totally impractical for all but the most modestly-sized problems; FFT methods offer an excellent strategy for perfectly stationary problems, however such stationarity is rare in remotely-sensed measurements; local methods compute

The author is with the Department of Systems Design Engineering, University of Waterloo, Canada, N2L-3G1.
Tel: (519) 888-4567 x3599 Fax: (519) 746-4791 Email: pfieguth@uwaterloo.ca
This research was supported in part by the Office of Naval Research under Grant N00019-91-J-1004, and by the Natural Sciences and Engineering Research Council of Canada.
EDICS # IP 2-WAVP

estimates from a local subset of measurements, which does not support data fusion with non-local measurements, and which may be undesirable for processes having long correlation lengths[2].

Instead, we consider a promising alternative approach to estimation which involves a recursive or hierarchical divide-and-conquer strategy[7, 14, 15]: a subset $\boldsymbol{x}_o$ of the random field is found such that conditioning on it, certain remaining portions of the field can be processed independently. In the context of estimation, this implies that the subset $\boldsymbol{x}_o$ conditionally decorrelates the remaining portions $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$:

$$E[\boldsymbol{x}_i \boldsymbol{x}_j | \boldsymbol{x}_o] = E[\boldsymbol{x}_i | \boldsymbol{x}_o] \cdot E[\boldsymbol{x}_j | \boldsymbol{x}_o]. \tag{1}$$

For example, the four quadrants of a first-order Markov random field[5, 6] can be decorrelated by conditioning on the boundary pixels shown in Figure 1. More generally, for first-order fields, a single pixel can decorrelate two halves of a one-dimensional process, a *column* of pixels is required for a 2D field, and a whole *plane* of pixels in three dimensions. This process of conditional decorrelation can be continued recursively, which gives rise to a tree structure (as sketched in Figure 5 for a two-dimensional process).

However the need to conditionally decorrelate a large problem into separate pieces is also the fundamental weakness of divide-and-conquer strategies: the conditional decorrelation step is not possible for problems of arbitrary size (Figure 4), due to issues of computational complexity and numerical stability. We emphasize that these issues will arise for *all* (except pathologically-trivial) prior statistical models: the statistical degrees of freedom represented by the boundary must be at least proportional to the number of pixels, $(n^{(d-1)})$. Therefore the dimensional of the decorrelating state $\boldsymbol{x}_o$ will grow as $\mathcal{O}(n^{(d-1)})$, and so matrix conditioning and computational complexity issues, or a corresponding sacrifice in the degree of statistical decorrelation, are asymptotically unavoidable. The whole purpose of this paper is to develop an alternative statistical framework with improved asymptotic properties by avoiding the decorrelation of arbitrarily-large regions. We begin by discussing the three central issues: computational complexity, numerical stability, and statistical accuracy.

Using divide-and-conquer, the computational effort to solve an estimation problem over an $n \times n \times \ldots$ hypercube of voxels in d dimensions is dominated by the estimation problem at the coarsest state $\boldsymbol{x}_o$; the effort

to invert the error covariance matrix associated with $\boldsymbol{x}_o$ is

$$\mathcal{O}\left(\left\{n^{(d-1)}\right\}^3\right) \tag{2}$$

$$= \quad \mathcal{O}(n^{(2d-3)}) \text{ per pixel} \tag{3}$$

which, although vastly less than the $\mathcal{O}(n^{3d})$ required for brute-force inversion, still becomes infeasible for large n and even modest d, necessitating some degree of order reduction.

Although much of this paper will be motivated by the computational complexity issues raised in the previous paragraph, an equally serious issue is that of numerical stability. The $q \times q$, $q \sim \mathcal{O}(n^{d-1})$, covariance matrix characterizing the statistics of the coarsest state has a condition number that grows with q. In many cases, machine precision, rather than computational power, may determine the upper limit on feasible n, d. Indeed, Table 1 plots the condition number of the coarsest state for a sample problem in $d = 2$ dimensions as a function of size n and sampling density ζ. Blank entries in the table correspond to problems which are numerically unstable (i.e., yielding negative-definite covariances) despite using double-precision floating-point and a Joseph-stabilized form[4] of a multiscale estimator[7]. Clearly as the problem size n is increased, even at modest sizes we are increasingly forced to reduce ζ, that is, to compromise statistical fidelity.

Statistical fidelity refers to the extent to which a model is an approximation of the true statistics; that is, the extent to which (1) holds true for a particular choice of state $\boldsymbol{x}_o$. An approximate model can affect the reliability of the estimates and estimation error statistics; however in many circumstances a greater concern revolves around the introduction of estimation artifacts at the hierarchical boundaries[18, 19, 22], due to an inadequate decorrelation in (1). For example, the two pixels $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ in Figure 2 are neighboring in physical space, but are distantly separated in terms of the hierarchical model, since their common parent is $\boldsymbol{x}_o$, all the way up the tree at the coarsest scale.

In this paper we address the above three issues by taking a new approach: rather than requiring the statistical division of a huge problem into pieces, we instead attack the problem one small piece at a time, limiting the size of regions needing to be decorrelated, and therefore avoiding asymptotic computational and numerical

problems. Our approach is *not*, however, a local one: each small piece of the problem is estimated based on *all* of the measurements, and the statistical model used is capable of representing non-local measurements and of performing data-fusion. Section 2 motivates this alternative framework, followed by its multiscale implementation, and finally introducing multiply-rooted trees as a computationally attractive means for realizing the model. Section 3 discusses the application of multiply-rooted trees to two large problems in remote-sensing: ocean-surface temperature and altimetry (height) estimation. The paper is concluded in Section 4.

2 Hierarchical Estimation

A. Motivation

In this paper will use as our context a multiscale statistical estimation framework[7, 10, 22], although the discussion and concepts presented in this paper apply equally to other hierarchical methods, such as nested-dissection[14, 15]. The principle (1) of hierarchical or recursive divide-and-conquer, as applied to a random field, is as follows: each state in the hierarchy must conditionally decorrelate its immediate descendents from each other and from the rest of the domain; in the multiscale framework, this principle is asserted via the following statistical model:

$$x(s) = A(s)x(s\bar{\gamma}) + B(s)w(s), \tag{4}$$

where s is an index on a tree with parent $s\bar{\gamma}$, A and B are deterministic matrices, and w is a white-noise process, independent of $x(0)$ at the unique root of the tree. The whiteness of w implies that the state $x(s\bar{\gamma})$ must conditionally decorrelate all states connected to $s\bar{\gamma}$.

The state $x(s)$ is a vector-valued process on a tree, such as the two-dimensional quad-tree sketched in Figure 5, although in principle any tree (acyclic graph) can be accommodated. Noisy observations

$$y(s) = C(s)x(s) + v(s) \tag{5}$$

are available at each node, where $C(s)$ encodes the measurement model, and $v(s)$ is the white measurement

noise. Given the model $A(s), B(s), C(s)$, the measurements $y(s)$, and the prior covariance $P(0)$ at the root node, a two-sweep algorithm, modeled on the Rauch-Tung-Striebel smoother[25], has been developed[7, 21] which computes the estimates $\hat{x}(s)$ and estimation error covariances $\tilde{P}(s)$. The estimation algorithm itself is exact; the estimates are approximate only to the extent that the model (4), (5) itself is approximate.

In the case of first-order Markov random fields, condition (4) is easily satisfied, although at great computational cost, by letting $x(s\bar{\gamma})$ densely sample the boundaries of its four descendents, as shown in Figure 1. To be sure, a variety of state reductions have been proposed for Markov-like random fields in the multiscale setting:

1. Allowing neighboring regions to overlap[19] to reduce the effect of artifacts caused by residual state-to-state correlations;

2. Subsampling the pixels along the state boundary[23], as illustrated in Figure 3;

3. Taking averages or wavelet transforms of the boundary pixels[21];

4. Determining from the statistics of the boundary pixels the optimum linear functionals[18, 19] which maximize the decorrelation.

However none of these methods change the *asymptotic* behavior of the computational complexity, since the state dimension at the root of the decomposition is, in each case, $\mathcal{O}(n^{(d-1)})$. That is, for each of the above methods the root state dimension is *still* proportional to the number of boundary pixels, albeit with smaller multiplicative constants. Therefore these methods are effective at making practical the solution of ever larger estimation problems, however extremely large (global scale) and three-dimensional problems remain out of reach.

The problem with each of the above four methods (in particular, the state subsampling approach of Figure 3) is that their development was motivated by approximating the exact approach of Figure 1. However it is not *remotely* obvious that an approximation to the exact solution ((1), Figure 1) represents the best (or even an adequate) approach to approximate estimation. Indeed, arguably the state in Figure 3 attempts to decorrelate too much: in terms of estimating the top-left quadrant, keeping the details of the distant part of the bottom-right

quadrant is largely irrelevant; that is, the reduced state of Figure 7 will perform very nearly as well, even though it makes no attempt to satisfy (1). Although we now require four such reduced models (one for each quadrant), the state dimension needs to be reduced by only a factor of $4^{1/3}$ for the total computational effort to equal that of the earlier reduced-order state (Figure 3).

The principle can be illustrated using Markov random fi elds[5, 6]. Figure 8 shows the estimates produced by a multiscale tree, having as its root node a state model based on Figure 7. We can compute four such sets of estimates, one for each quadrant, each computed by a separate multiscale model and tree; the estimates were mosaiced together to produce the estimated fi eld shown in Figure 9, albeit at about one half the effort of estimating the entire domain directly, shown in Figure 6. Observe that despite the reduced computational effort, the typical artifacts at the quadrant boundaries (Figure 6) are not present. This example is an illustration only, in that other methods[10, 18, 23] have also been developed to successfully deal with these artifacts; we will argue that this sort of multiple-model approach addresses the computational and statistical issues as well.

Although the preceding example seems promising, we have not, in fact, changed the asymptotic complexity, since solving each of the quadrants represents solving one fourth of the whole domain; that is,

$$\text{Complexity}(\text{Quadrant}) \geq \frac{1}{64}\text{Complexity}(\text{Whole}) \tag{6}$$

$$\mathcal{O}(\text{Quadrant}) = \mathcal{O}(\text{Whole}) = \mathcal{O}\left(n^{(d-1)^3}\right) \tag{7}$$

Really, Figure 7 represents only the fi rst step in model subdivision; in general, we can choose to create p trees, each responsible for $1/p$ of the original domain. The key to changing the asymptotic complexity, and the key novel aspect of this paper, lies in ever further decomposing the tree as the size of the problem is increased. We can begin by considering only the complexity of the p root nodes, in which case the total effort *per pixel* goes as

$$\mathcal{O}\left[p \cdot \left(\frac{n}{\sqrt[d]{p}}\right)^{(d-1)^3} \cdot n^{-d}\right] = \mathcal{O}[p^{3/d-2} \cdot n^{2d-3}] \tag{8}$$

from which it should be noted that the benefi ts of this approach become more pronounced in higher dimensions,

as d increases.

The above model is a little simplistic in its assessment of the state dimension of each root. Figure 10 presents a more realistic example, sketched for the two-dimensional case, in which we estimate f^d pixels **A**, based on a detailed representation of the surrounding $(f+g)^d$ pixel area **B**, and an approximate representation of the rest of the domain **C**, where g is chosen based on the correlation length of the process statistics. The computational effort, per-pixel, of the model root nodes is approximately

$$\begin{aligned} & \mathcal{O}\left[(\#\text{ Models}) \cdot (\text{Effort per Model})/(\#\text{ Pixels})\right] \\ = \; & \mathcal{O}\left[\left(\frac{n^d}{f^d}\right)\left[\left(\xi(f+g)^{d-1}\right)^3\right]/n^d\right] \qquad (9) \\ = \; & \mathcal{O}\left[\frac{\xi^3(f+g)^{3d-3}}{f^d}\right] \qquad (10) \end{aligned}$$

which is minimized by setting

$$f = \frac{dg}{2d-3}, \qquad (11)$$

that is, typically the optimum size will be a fixed value $f \ll n$. Substituting back into (10) we find that, for a fixed dimension d, the computational effort per-pixel goes as

$$\mathcal{O}\left(g^{2d-3}\right). \qquad (12)$$

That is, the complexity per pixel is strictly a function of correlation length, as is intuitive, and is *not* a function of n, in sharp contrast to the original multiscale model (3).

Although the effort (10) appears to ignore the fact that each tree has n^d pixels on the finest scale which must be estimated, suggesting an effort more fairly represented as

$$\frac{\xi^3(f+g)^{3d-3}}{f^d} + \beta\frac{n}{f}^d, \qquad (13)$$

almost all of the processing at the finest scales is the same for all p models, as discussed below, so (10) is indeed

a realistic reflection of the algorithm proposed in this paper.

B. Implementation

Our proposed model is sketched in Figure 11, in which the region to be estimated is embedded in a nested hierarchy, represented in decreasing statistical fidelity. We have chosen to implement the model in the context of a recent-developed multiscale estimation framework[1, 7, 21], which leads to the following advantages:

- The existence of efficient estimation[7] and likelihood[22] algorithms.
- An existing base of multiscale models to represent the statistics of the process being modeled.
- A stochastic realization theory[19].
- The ability to represent both local and non-local measurements.
- Desirable asymptotic properties, based on the proposed model of this paper.

We need to define p models; to make the model structure as regular as possible, we will select some scale $\bar{m}$ and develop $p = 2^{d\bar{m}}$ models, such that model $\mathcal{M}_i$ represented on tree $\mathcal{T}_i$ estimates the region associated with the ith multiscale state on scale $\bar{m}$. Let this ith state be defined as tree node t_i. Each of the p trees is still a model of the entire domain, in that the finest scale on each tree has n^d states (one per pixel), although clearly the quality of the model will vary from state to state on a given tree. The scale $\bar{m}$ is chosen to give sensible values for f, g:

$$f = g/2 = n \cdot 2^{-\bar{m}}, \tag{14}$$

based on the discussion in the previous section.

Each model $\mathcal{M}_i$ is uniquely defined by the statistics of the process and the nature of the state $x(s)$[19, 23]. We propose to define the state at tree node s as a subsampling of the union of the boundaries of the children of node s. The state sampling density $\zeta(s) \in \{\zeta_A, \zeta_B, \zeta_C\}$, which measures the number of state samples along the child boundaries per correlation length, affects the statistical fidelity to which $x(s)$ is modeled, and it is the

differences in statistical fidelity by which the models $\mathcal{M}_i$ differ. Specifically, the fidelity follows the structure of Figure 12, that is, a decreasing function of the spatial separation between $x(s)$ and the region to be estimated. Each tree $\mathcal{T}_i$ is built in three stages:

1. At scale $\bar{m}$, the state $x(s)$ is based on the sampling density $\zeta(s)$, which follows from Figure 11.

2. At scales finer than $\bar{m}$ the tree is regular with 2^d descendents per node, and where $\zeta(s) = \zeta(\sigma)$, where σ is the ancestor of s on scale $\bar{m}$.

3. At scales coarser than m the tree is structurally simple, but awkward to explain. The structure is built upwards from scale $\bar{m}$, following the example of Figure 13:

   - Node t_i and each of its ancestors represent the spatial region to be estimated (albeit with varying statistical fidelity).
   - Parent node $t_i\bar{\gamma}$ has up to 3^d descendents — the spatially closest neighbors of t_i, modeled at high precision ζ_A.
   - Grandparent node $t_i\bar{\gamma}\bar{\gamma}$ has up to 3^d descendents — the spatially next-closest neighbors of t_i, modeled to moderate precision ζ_B.
   - The remainder of the tree is regular, modeled at precision ζ_C, except where preempted by the above two points.

Figure 12 shows an example for the one-dimensional case; one of $p = 8$ models are shown, so the nested structure of Figure 11 appears on scale $\bar{m} = \log_{2^d}(p) = 3$. Coarser states are found by grouping those on scale 3; on all scales finer than scale 3 the tree is just regular dyadic. An analogous example for the two-dimensional case for one of $p = 64$ models is shown in Figure 13; beyond scale 3 the tree is a regular quadtree.

It is the above spatial arrangement of descendents, rather than a rigid adherence to a quadtree structure, and the fact that each root node represents only a small fraction (one *pth*) rather than the whole domain, which is the essence of our approach and leads to improved numerical conditioning and pixel correlations. Given estimates $\hat{x}(s)$ from each tree $\mathcal{T}$, to produce estimates for the whole process we mosaic the estimates associated with

each region: that is, take $\hat{x}(s)$ from tree $\mathcal{T}_i$ whenever s is a descendent of s_i, for each of the $1 \leq i \leq p$ models. The usual multiscale model [7, 21, 22] is used to solve for the estimates on each tree; the p estimation runs for the p trees are not coupled and can be executed in parallel.

Certain practical aspects of the algorithm remain to be addressed in the next section, however some important numerical and statistical properties can be assessed at this point. As discussed in the introduction, existing divide-and-conquer approaches are disadvantaged in solving large problems for three reasons: computational effort, numerical conditioning, and artifacts due to statistical approximations. The proposed algorithm addresses each of these:

1. The computational effort is reduced, as discussed, by creating a set of models, each model responsible for only a small portion of the process. We never attempt to conditionally decorrelate quadrants of the entire process.

2. The numerical conditioning is improved by avoiding states with high dimensions. As shown in Table 1, as the size of a regular multiscale process increases, ever poorer values of ζ must be accepted to keep the algorithm stable, whereas the conditioning of the proposed algorithm (last row of Table 1) is essentially independent of scale.

3. The fidelity of the statistical model related to the sampling density ζ. Table 2 plots the correlation-coefficient ratios which we use as a measure of model statistical fidelity:

$$\frac{1.0 - \rho_{REALIZED}(s_1, s_2)}{1.0 - \rho_{TRUE}(s_1, s_2)}. \tag{15}$$

That is, we measure the ratio of the deviations of the true correlation coefficient ρ and the multiscale-realized coefficient from one, for two pixels s_1, s_2 separated by a major tree boundary (as in Figure 2). The limited range of permissible ζ, imposed by the numerical-stability limits of Table 1, increasingly limits the multiscale model, with corresponding increases in statistical inconsistencies (15) and estimation artifacts as the size of the problem increases. The multiple-model alternative is not subject to the same

constraints on ζ, so ζ can be chosen (last row of Table 2) to meet statistical objectives for problems of any size.

C. Multiply-Rooted Trees

The proposed method uses the multiscale approach to develop p separate models on p separate trees, and then mosaics a subset of the estimates from each tree to produce a set of estimates for the entire random process.

The most obvious criticism of using p completely separate models to estimate a random field is that such an approach ignores the fact that the models may involve a great deal of duplicated effort. For example, the estimation of state α (on scale 3 of Figure 13) will be required for nine different models; even greater duplication occurs on finer scales, where the redundancy will be as large as p for many nodes. In fact, by detecting and removing such duplication, the memory to represent the union of *all* required nodes on all p models simultaneously is comparable to that required for traditional, single-tree, methods: although we now need to represent some nodes multiple times (i.e., at various samplings ζ_i), we have corresponding savings due to the absence of coarse-scale nodes with large state dimensions (and correspondingly huge covariances). Our goal is to derive a graphical structure representing the *union* of the p tree models, which removes the duplication present in the models, produces exactly the same estimates, and which leads to a computationally-efficient algorithm, similar to that which exist for normal multiscale trees. We will show that a single model, with p roots, can achieve these goals. Note that the resulting model, although no longer a tree, is an efficient concatenation of the p multiscale tree models, and so the usual fast tree-based algorithm[7, 21] still applies. The model of this paper is thus in sharp contrast to estimation methods such as [9, 20] on graphs, rather than trees, and which are computationally much more demanding.

Let $\mathcal{T}_i$ be the tree corresponding to the ith model, and let $\mathcal{T}_i(s)$ be the subtree of descendents from node s. We define Y to be all available measurements, and $Y(\mathcal{T})$ to be the measurements on tree $\mathcal{T}$. Each node s on a tree is characterized in terms of its parent $s\bar{\gamma}$, its children $s\alpha_i$, its sampling density $\zeta(s)$, its scale $m(s)$, and its spatial location/extent $\pi(s)$. On a given tree $\mathcal{T}$, the multiscale estimator[1, 7] proceeds in three stages:

1. An *Upwards* pass, in which the conditional estimates at node s

$$\hat{x}_U(s) = E[x(s)|Y(\mathcal{T}(s))] \tag{16}$$

are computed based on the measurements in the subtree below s.

2. At the root of the tree,

$$\hat{x}_D(0) = \hat{x}_U(0). \tag{17}$$

3. A *Downwards* pass, in which the smoothed estimates

$$\hat{x}_D(s) = E[x(s)|Y] \tag{18}$$

are found. These are the *best* linear estimates based on all of the measurements on the whole tree, but computed locally as

$$\hat{x}_D(s) = f(\hat{x}_U(s), \hat{x}_D(s\bar{\gamma})). \tag{19}$$

For model $\mathcal{M}_i$ on tree $\mathcal{T}_i$, we are interested only in the estimates within $\mathcal{T}_i(t_i)$; that is, the set of estimates from $\mathcal{T}_i$ which we wish to mosaic. Therefore the downwards pass is relevant only on $\mathcal{T}_i(t_i)$ and on $\left\{t_i\bar{\gamma}^k\right\}$, the ancestors of t_i.

Let

$$\mathcal{T}_* = \mathcal{T}_1 \bigcup \ldots \bigcup \mathcal{T}_p \tag{20}$$

be the union of the p models, as sketched in Figure 14. We define the equivalence of two nodes $s_i \in \mathcal{T}_i, s_j \in \mathcal{T}_j$

as

$$s_i \equiv s_j \iff \zeta(s_i) = \zeta(s_j), \pi(s_i) = \pi(s_j), m(s_i) = m(s_j) \tag{21}$$

The estimates on $\mathcal{T}_*$ are equal to the mosaic of estimates from the p separate models if the following conditions hold:

1. Upwards Pass: For any nodes $s_i \in \mathcal{T}_i, s_j \in \mathcal{T}_j$ on the same scale, the construction of the p trees guarantees that

$$s_i \equiv s_j \Longrightarrow \mathcal{T}_i(s) \equiv \mathcal{T}_j(s) \tag{22}$$

   and therefore that

$$s_i \equiv s_j \Longrightarrow \hat{\boldsymbol{x}}_U(s_i) = \hat{\boldsymbol{x}}_U(s_j) \tag{23}$$

   That is, all of the estimates from the p upwards passes will be present on combined tree $\mathcal{T}_*$.

2. Let the p roots be $0_1, \ldots, 0_p$. From before, by definition $\hat{\boldsymbol{x}}_U(0)$ on tree $\mathcal{T}_i$ equals $\hat{\boldsymbol{x}}_U(0_i)$, therefore the downwards pass is initialized as

$$\hat{\boldsymbol{x}}_D(0_i) = \hat{\boldsymbol{x}}_U(0_i), \tag{24}$$

   exactly as in (17).

3. Downwards Pass: The downwards pass on each tree $\mathcal{T}_i$ traverses the ancestors (scales $\leq \bar{m}$) and descendents (scales $> \bar{m}$) of node t_i, dividing the pass into two parts.

   On scales $\leq \bar{m}$, the downwards pass on tree $\mathcal{T}_i$ proceeds only between those ancestors $t_i\bar{\gamma}^k$ of node t_i representing the spatial region of interest; that is, for which $\pi(t_i\bar{\gamma}^k) = \pi(t_i)$. The situation is complicated

on $\mathcal{T}_*$ because a node s may have multiple parents (that is, our structure is no longer a tree), whereas (19) implicitly assumed unique parentage. The desired parent for the downwards pass is the one with the same region, that is, (19) is modified as

$$\hat{x}_{D}(s) = f(\hat{x}_{U}(s), \hat{x}_{D}(s')) \ \ni \ m(s') = m(s) - 1, \ \pi(s') = \pi(s) \tag{25}$$

On scales $> \bar{m}$, $\mathcal{T}_*$ is a regular tree, and the downwards pass of (19) remains unchanged.

3 Experimental Results

We have already demonstrated the most fundamental results of multiply-rooted trees in the context of numerical conditioning in Table 1 and statistical fidelity in Table 2. Of the three criteria listed in the Introduction, the remaining one is computational complexity. Our research goal is the development of statistical methods for very large problems, so in this section we focus on computational issues.

Table 3 shows the improvement in speed of our proposed approach over the standard singly-rooted multiscale algorithm[7, 22], when applied to the Markov random field problem of Figure 9. The extra states introduced by our multiply-rooted approach cannot be justified for extremely small or poorly-sampled problems (upper left of Table 3), however as the problem size and sampling density increase (lower right) the decomposition offered by the multiply-rooted approach becomes more competitive. For large, densely sampled trees, computational improvements in excess of a factor of twenty were observed.

We further demonstrate the application of multiply-rooted trees to two challenging remote sensing problems: ocean altimetry (height) [10, 27] and ocean surface-temperature estimation[23, 24]. Both problems are of substantial scientific interest and represent aspects of ongoing collaborative efforts. More significantly, these two estimation problems are the ones which originally motivated the research of this paper through the possession of the following attributes:

1. A high resolution is required to preserve features of interest (e.g., ocean eddies).

2. The desired size of the solution is enormous (basin-scale or global-scale).

3. The statistics of the problems are challenging, coupling extremely accurate measurements with a prior model having very large variances.

A. Ocean Altimetry

Ocean altimetry data are detailed measurements of the height of the ocean surface, collected by radar via satellite (in this case, the joint American/French TOPEX/POSEIDON satellite[27]) to astounding accuracies — about 5cm height error standard deviation, from a satellite orbiting at over 1000km in altitude. The altimetry data contains information relevant to geodesy (the shape of the ocean surface reflects variations in the earth's gravitational field) and oceanography (the presence of ocean currents induces, via the Coriolis effect, a slope in the height of the ocean surface).

We have investigated the use of multiscale models for altimetry data in earlier studies[10, 11], however those studies employed a much simpler scalar model

$$x(s) = 1 \cdot x(s\bar{\gamma}) + B2^{(1-\mu)m(s)/2}w(s) \tag{26}$$

in which a single scalar $x(s)$ is relied upon to decorrelate the four children of s, which would be inadequate for problems of large size.

Figures 15 through 17 present the altimetric results. Figure 15 shows a set of 21000 sparse altimetric measurements, distributed over the north Pacific. Figures 16 and 17 show the 200×200 estimates and estimation error results, based on a prior model having a correlation length of 15 pixels (3 degrees) and a standard deviation of 100cm. The results are taken from a 64-root, 9-scale overlapped[18] tree with a state sampling every four pixels (a model which would be numerically unstable given the traditional, single-root estimator). The effort to compute the estimates and error variances is about 105 seconds on a Sun ULTRA-1 (actual wall-clock time is several minutes longer due to code overhead).

B. Ocean Temperature

An enormous amount of ocean surface-temperature data is collected using passive radiometers on a variety of satellite platforms. Among the best of these sensors is the Along Track Scanning Radiometer (ATSR)[24], mounted on European Space Agency's ERS-1/2. The detailed examination of ocean-surface temperature maps is important for general oceanographic studies, such as the shape of the Gulf stream, or for studies of climate change, assessing heat fluxes and long-term temperature trends.

The infrared emission strength of the ocean surface is observed at several wavelengths and through two atmospheric paths, all of which is converted into a temperature measurement, such as those shown in Figure 18, via a preprocessing stage. In this collaboration[13], interest is focussed on the short-term dynamics of the ocean temperature, so the data is mean-removed (i.e., the static or systematic component is subtracted).

Temperature estimation via multiscale means (using a single tree) was examined in an earlier study[23], however that paper was oceanographic-scientifi c in nature rather than computational or algorithmic. The model employed in that study was a reduced-order one, similar to that sketched in Figure 3, however the size of the state dimension was introducing numerical challenges, such that the processing of larger domains would have been very diffi cult.

Figures 19 and 20 show the temperature estimates and estimation error statistics for a 200×200 pixel region in the equatorial Pacifi c, based on a prior model having a correlation length of 30 pixels (5 degrees) and a standard deviation of 3K; the 5000 measurements have an error standard deviation of 0.15K. The results are computed from a 64-root, 9-scale overlapped[18] tree with a state sampling every ten pixels — that is, 3.0 state elements per correlation length. By contrast, our earlier reduced-order approach reported in [23] was limited to between 1.0 and 2.3 state elements per correlation length for reasons of numerical stability.

The effort to compute the estimates and error variances is about 40 seconds on a Sun ULTRA-1 (total wall-clock time of about two minutes).

4 Conclusions

We have presented a new approach to the estimation of large random fi elds, based on a recently-introduced class of multiscale stochastic processes. The work was motivated by the growing class of huge estimation problems, many in the area of terrestrial remote sensing, and by the observation that divide-and-conquer approaches to estimation (whether multiscale, nested-dissection, or otherwise) eventually run into numerical diffi culties for asymptotically large estimation problems.

This paper has introduced a new multiscale estimation formulation, one which should be applicable to estimation problems of extremely large size. A variety of issues remain unanswered — the proper size and geometry of regions in order to "adequately" estimate the local region of interest, the choice of states, and appropriate ways in which to mosaic the individual estimates. However there is ample motivation to pursue further research into the aforementioned diffi culties of the proposed formulation, both because of its desirable asymptotic complexity, but also because it preserves the desirable properties of the multiscale estimation framework[1, 7, 10, 21] from which it is derived: the existence of effi cient estimation and likelihood algorithms, an existing base of multiscale models, the ability to represent local and non-local measurements, and a stochastic realization theory.

Acknowledgements

We would like to thank the anonymous reviewers for helpful suggestions and insights to clarify the presentation of the paper.

References

[1] M. Basseville, A. Benveniste, K. Chou, S. Golden, R. Nikoukhah, A. Willsky, "Modeling and estimations of multiresolution stochastic processes.", *IEEE Transactions on Information Theory* (38) #2, pp.766–784, 1992

[2] J. Beran, *Statistics for Long-Memory Processes*, Chapman & Hall, 1994.

[3] C. A. Bouman and M. Shapiro, "A Multiscale Random Field Model for Bayesian Image Segmentation," *IEEE Trans. on Image Processing*, (3) #2, pp.162–177, March 1994.

[4] R. Bucy, P. Joseph, *Filtering for Stochastic Processes*, Wiley, 1968.

[5] R. Chellappa and S. Chatterjee. "Classification of textures using Gaussian Markov random fields." *IEEE Transactions on ASSP*, (33), pp. 959-963, 1985.

[6] R. Chellappa, A. Jain ed.s, *Markov Random Fields – Theory and Application*, Academic Press, 1993

[7] K. Chou, A. Willsky, A. Benveniste, "Multiscale Recursive Estimation, Data Fusion, and Regularization", *IEEE Trans. on Automatic Control* (39) #3, pp.464–478, 1994

[8] Chi-hsin Wu, Peter C. Doerschuk, "Tree Approximations to Markov Random Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (17) #4, pp.391–402, April 1995

[9] E. Fabre, "New fast smoothers for multiscale systems," *IEEE Trans. Signal Processing* (44) #8, pp.1893–1911, 1996

[10] P. Fieguth, W. Karl, A. Willsky, C. Wunsch, "Multiresolution Optimal Interpolation and Statistical Analysis of TOPEX/POSEIDON Satellite Altimetry," *IEEE Trans. Geoscience and Remote Sensing* (33) #2, pp.280–292, 1995

[11] P. Fieguth, D. Menemenlis, T. Ho, A. Willsky, C. Wunsch, "Mapping Mediterranean Altimeter Data with a Multiresolution Optimal Interpolation Algorithm", *Journal of Atmospheric and Oceanic Technology* (15), pp.535–546, 1998

[12] P. Fieguth, W. Karl, A. Willsky, "Efficient Multiresolution Counterparts to Variational Methods for Surface Reconstruction," *Computer Vision & Image Understanding* (70) #2, pp.157–176, 1998

[13] P. Fieguth, F. Khellah, M. Murray, M. Allen, "Large Scale Dynamic Estimation of Ocean Surface Temperature," *IGARSS'99*, Hamburg, 1999

[14] A. George, *Computer solution of large sparse positive definite systems*, Prentice-Hall, 1981.

[15] G. Golub, R. Plemmons, "Large-Scale Geodetic Least-Squares Adjustment by Dissection and Orthogonal Decomposition," in *Large Scale Matrix Problems* (Björk, Plemmons, Schneider eds.), North Holland, New York, pp.3–28, 1981

[16] L. Greengard, V. Rokhlin, "A Fast Algorithm for Particle Simulations," *J. Computational Physics* (73), pp.325–348, 1987

[17] W. Irving, W. Karl, A. Willsky, "A Theory for Multiscale Stochastic Realization", *33rd Conference on Decision and Control*, 1994.

[18] W. Irving, P. Fieguth, A. Willsky, "An Overlapping Tree Approach to Multiscale Stochastic Modeling and Estimation", *IEEE Trans. on Image Processing* (6) #11, pp.1517–1529, 1997

[19] W. Irving, A. Willsky, "A Canonical Correlations Approach to Multiscale Stochastic Realization," To appear in *IEEE Transactions on Automatic Control*

[20] J.M. Laferte, F. Heitz, P. Perez, E. Fabre, "Hierarchical statistical models for the fusion of multiresolution image data," *Proc. Int'l Conf. on Computer Vision*, pp.908–913, 1995

[21] M. Luettgen, W. Karl, A. Willsky, R. Tenney, "Multiscale Representations of Markov Random Fields", *IEEE Trans. Signal Processing* (41) #12, pp.3377–3396, 1993

[22] M. Luettgen, W. Karl, A. Willsky, "Effi cient Multiscale Regularization with Applications to the Computation of Optical Flow." *IEEE Transactions on Image Processing* (3) #1, pp. 41-64, 1994.

[23] D. Menemenlis, P. Fieguth, C. Wunsch, A. Willsky, "Adaptation of a Fast Optimal Interpolation Algorithm to the Mapping of Oceanographic Data", *Journal of Geophysical Research* (102) #C5, pp.10573–10584, 1997

[24] C.T. Mutlow and A.M. Zavody, “Sea surface Temperature Measurements by the along-track scanning radiometer on the ERS1 Satellite: Early results,” *Journal of Geophysical Research* (99) #C11, pp.22575-22588, 1994.

[25] H. Rauch, F. Tung, C. Striebel, “Maximum Likelihood Estimates of Linear Dynamic Systems”,*AIAA Journal*, (3) #8, 1965

[26] V. Rokhlin, “Rapid Solution of Integral Equations of Classical Potential Theory,” *J. of Computational Physics* (60), pp.187–207, 1983

[27] “Topex/Poseidon: Geophysical Evaluation,” *Journal of Geophysical Research* (99) #C12, 1994

# Scales	State Density ζ (# States per Correlation Length)								
$(1+\log_2 n)$	0.5	1.0	1.5	2.0	2.5	3.0	4.0	5.0	6.0
5	$2 \cdot 10^4$	$2 \cdot 10^4$	$2 \cdot 10^4$	$9 \cdot 10^4$	$3 \cdot 10^4$	$3 \cdot 10^4$	$9 \cdot 10^8$	$1 \cdot 10^9$	$2 \cdot 10^9$
6	$9 \cdot 10^2$	$4 \cdot 10^3$	$2 \cdot 10^3$	$8 \cdot 10^6$	$1 \cdot 10^7$	$9 \cdot 10^6$	$2 \cdot 10^9$	$4 \cdot 10^9$	$8 \cdot 10^{12}$
7	$2 \cdot 10^2$	$6 \cdot 10^4$	$4 \cdot 10^4$	$2 \cdot 10^6$	$4 \cdot 10^6$	$5 \cdot 10^9$	$3 \cdot 10^{10}$	$7 \cdot 10^{12}$	
8	$4 \cdot 10^2$	$4 \cdot 10^3$	$3 \cdot 10^6$	$4 \cdot 10^6$	$2 \cdot 10^8$	$5 \cdot 10^9$			
9	$3 \cdot 10^1$	$2 \cdot 10^3$	$7 \cdot 10^4$	$8 \cdot 10^6$	$2 \cdot 10^9$	$6 \cdot 10^{11}$			
10	$2 \cdot 10^1$	$4 \cdot 10^2$	$4 \cdot 10^4$	$2 \cdot 10^8$	$2 \cdot 10^{11}$				
11	$8 \cdot 10^0$	$4 \cdot 10^2$	$8 \cdot 10^4$	$3 \cdot 10^8$					
12	$8 \cdot 10^0$	$6 \cdot 10^2$	$3 \cdot 10^5$	$5 \cdot 10^8$					
Mult Root	$6 \cdot 10^1$	$8 \cdot 10^1$	$3 \cdot 10^3$	$5 \cdot 10^4$	$4 \cdot 10^4$	$6 \cdot 10^5$	$2 \cdot 10^7$	$1 \cdot 10^{11}$	$4 \cdot 10^{12}$

Table 1: Matrix condition number of the state having the largest dimension for a two-dimensional process with spatial Gaussian statistics, correlation length of 60 pixels. The condition number is plotted as a function of problem size ($n = 2^{\text{Scales}-1}$) and state density ζ; blank entries are numerically unstable. The bottom row shows the corresponding values for the proposed method, (nearly) independent of problem size.

# Scales	State Density ζ (# States per Correlation Length)								
$(1+\log_2 n)$	0.5	1.0	1.5	2.0	2.5	3.0	4.0	5.0	6.0
5	6	6	6	7	6	6	6	6	6
6	13	22	16	10	9	9	7	7	6
7	220	41	38	9	8	8	6	6	
8	490	75	50	10	9	9			
9	2000	97	36	9	8	8			
10	2000	120	33	9	8				
11	1900	120	33	8					
12	1600	110	31	8					
Mult Root	400	400	57	45	9	6	5	1	1

Table 2: Model statistical degradation (desired correlation / realized correlation) for two pixels straddling a coarse tree boundary (as in Figure 2). The degradation is plotted as a function of problem size ($n = 2^{\text{Scales}-1}$) and state density ζ; blank entries are numerically unstable. The bottom row shows the corresponding values for the proposed method, independent of size.

# Scales	State Density ζ				
$(1 + \log_2 n)$	0.5	1	1.4	3.5	5
6	0.2	0.4	0.7	3.7	4.0
7	0.4	1.0	2.0	7.7	8.5
8	0.6	1.8	3.6	15.4	17.6
9	0.9	3.2	6.8	29.5	35.3
10	1.4	5.9	12.7		
11	2.3				

Table 3: Reduction of computational effort of the proposed approach over the standard multiscale algorithm (measured via raw FLOP count). As expected, the benefit increases for more difficult problems: the reduction factor increases for larger trees and more finely sampled domains. The results are based on the tree-like prior of Figure 9.

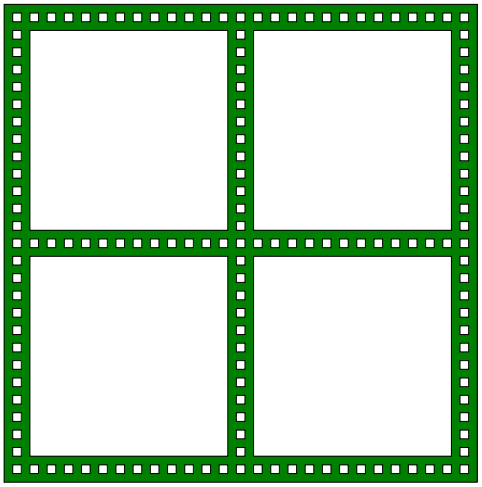

Figure 1: A dense set of boundary pixels, which would conditionally decorrelate the four quadrants of a first-order Markov random field.

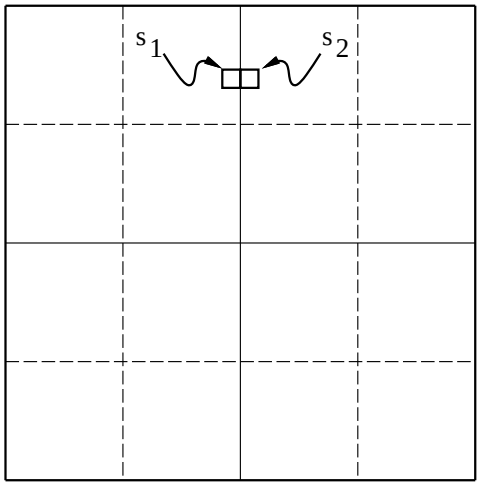

Figure 2: Two nodes, s_1 and s_2, neighbors in physical space, but distantly separated in tree space.

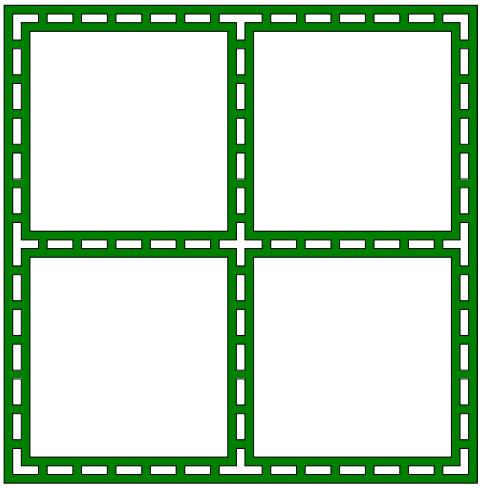

Figure 3: One possible reduced-order approximation to the state of Figure 1.

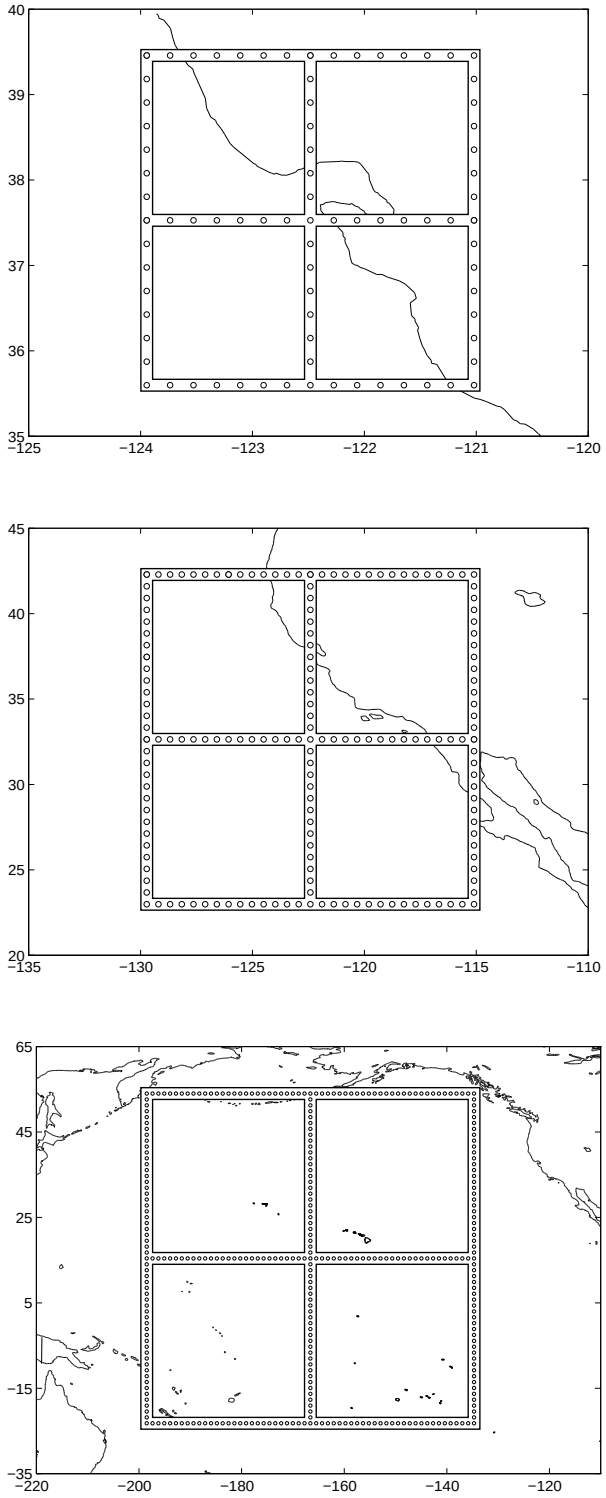

Figure 4: For how large a domain is divide-and-conquer computationally and numerically feasible . . . ?

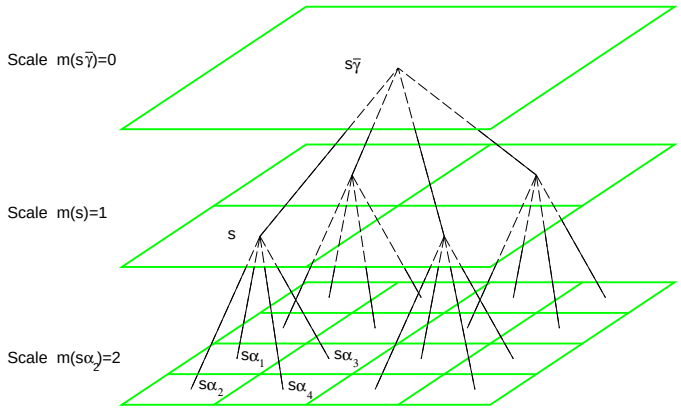

Figure 5: Illustration of the first three levels of a quad-tree.

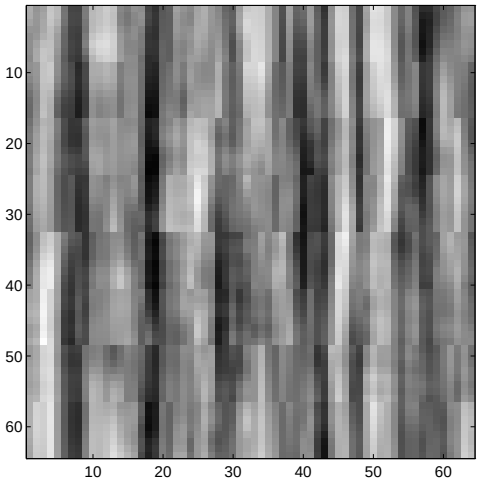

Figure 6: The estimation of a Markov random field using a multiscale tree based on a reduced-order model, as in Figure 3. Note the appearance of estimation artifacts at several tree boundaries.

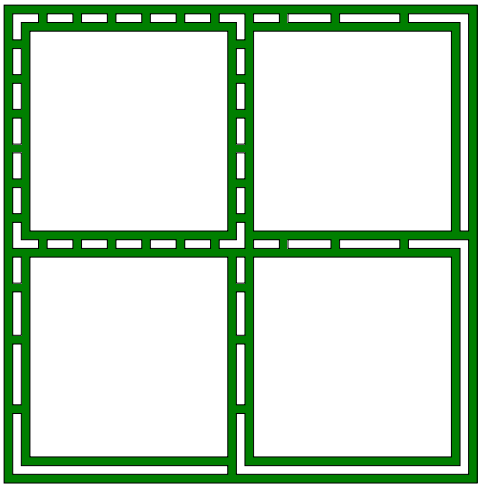

Figure 7: An alternative choice of reduced state, appropriate for estimating the upper-left quadrant of a process.

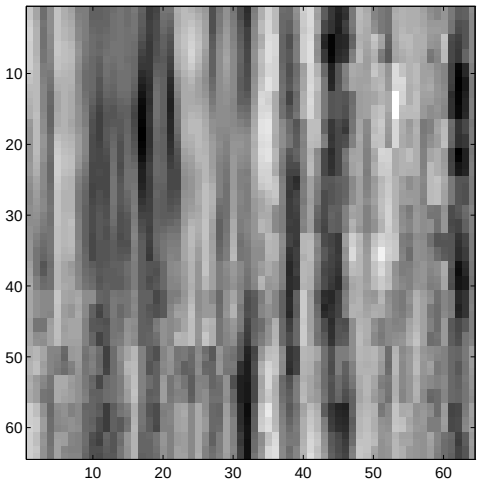

Figure 8: Estimates of a Markov random fi eld, based on a single-quadrant model (such as in Figure 7).

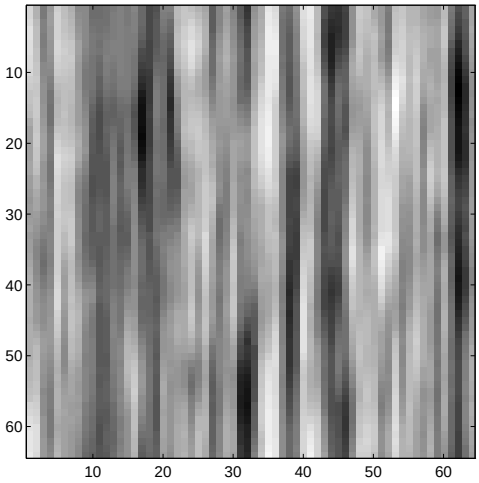

Figure 9: The mosaic of four sets of estimates, one set of estimates per quadrant as shown in Figure 8. Note the absence of artifacts.

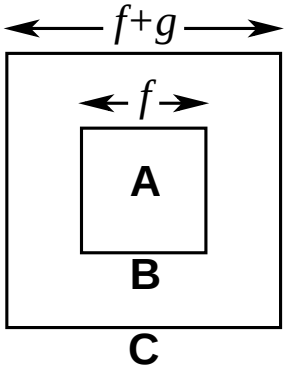

Figure 10: Proposed approach: construct a hierarchy of statistical representations, such that we estimate $f \times f$ region **A** based on a detailed representation of **B** and a more moderate one of **C**; g, controlling the width of **B**, is related to the correlation length of the process.

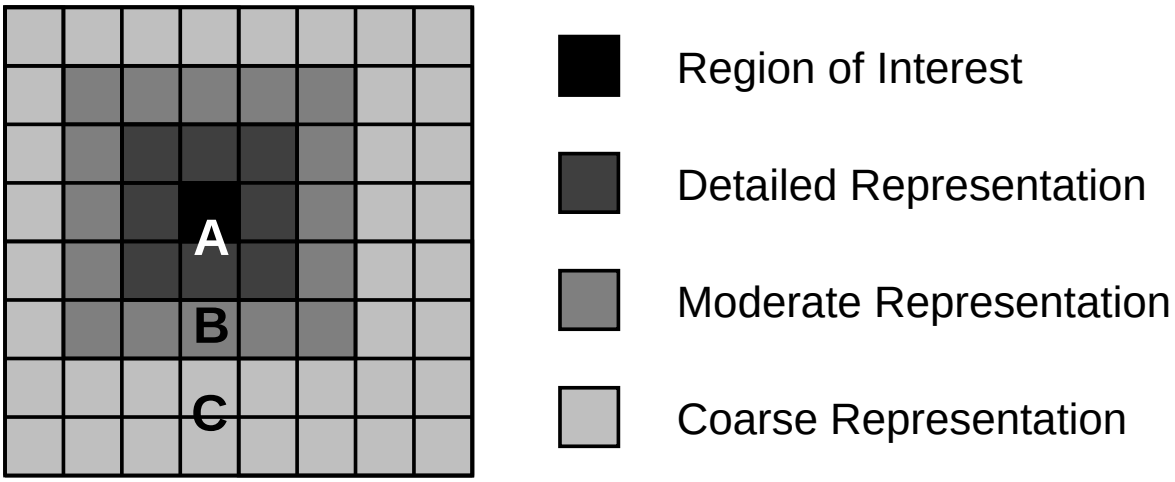

Figure 11: Specifi c approach: the problem is broken into $2^{d(\bar{m})}$ pieces on scale $\bar{m}$, where $\bar{m}$ is selected to keep the size $n2^{-\bar{m}}$ of each piece constant, independent of total problem size n. Each piece **A** is surrounded by two concentric bands of decreasing statistical detail ζ_A, ζ_B, followed by a coarse representation **C** of the remaining domain.

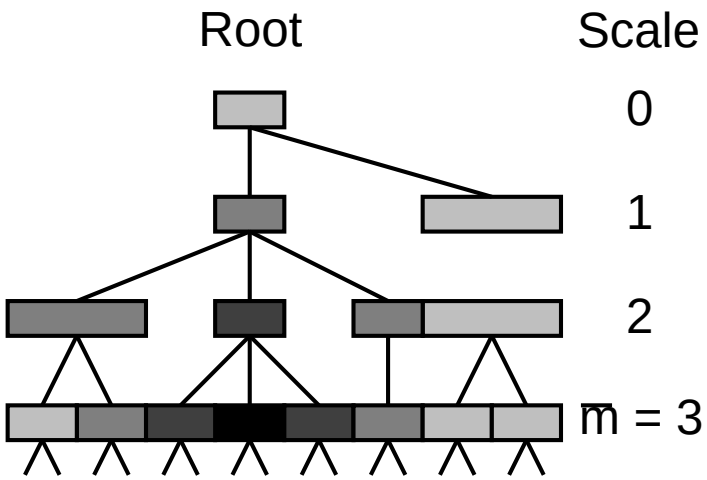

Figure 12: Example multiscale implementation for a 1D example; below the fourth scale the tree is regular dyadic.

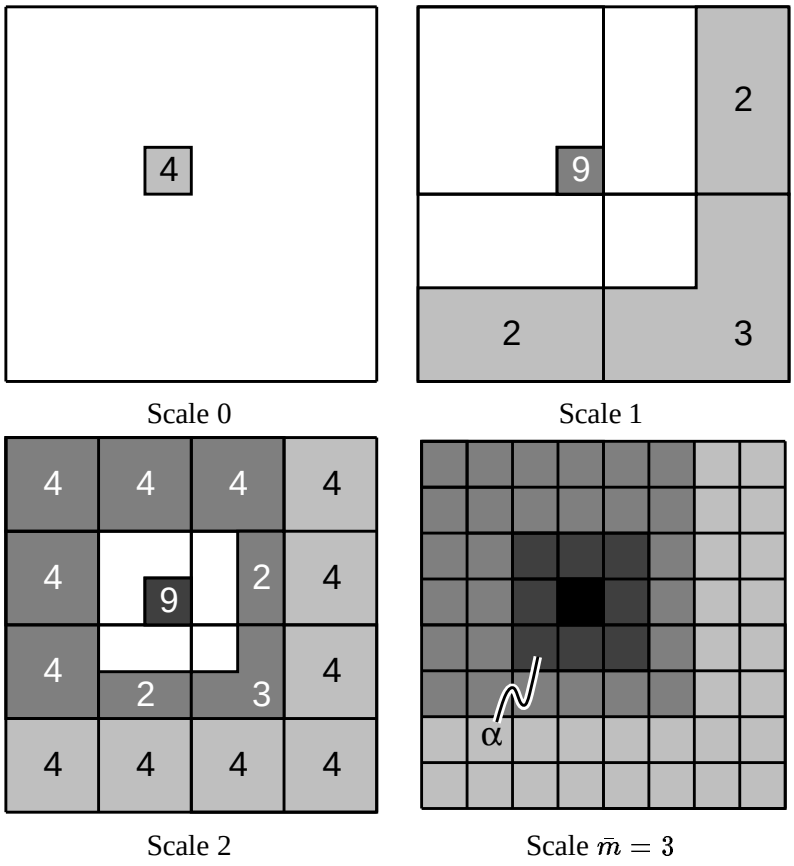

Figure 13: Example multiscale implementation for a 2D example; the structure is a regular quadtree below the fourth scale. The number within each node indicates the number of descendents. Note that this type of model is highly redundant; for example, the state at node α will be used in nine different models.

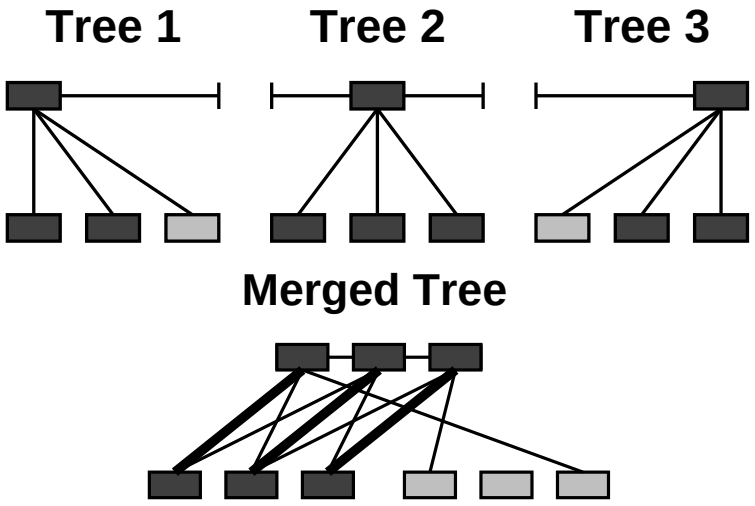

Figure 14: The merging of three trees; all arcs are followed for the upwards pass, only the thick arcs are followed downwards.

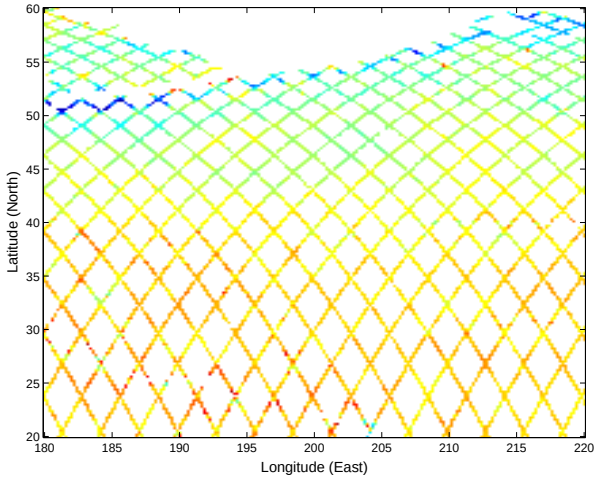

Figure 15: Measurements of ocean height in the northern Pacifi c.

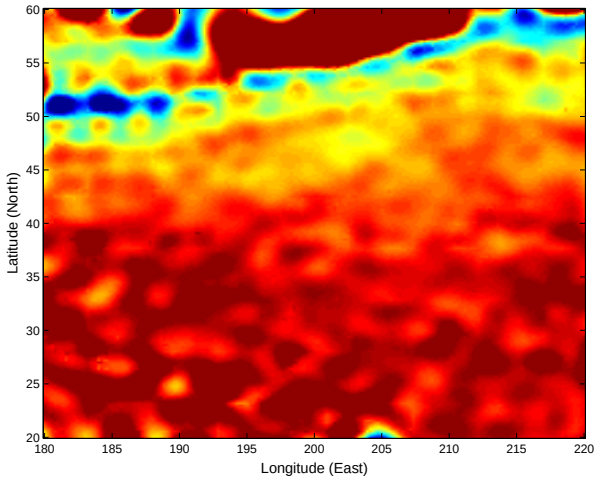

Figure 16: Dense estimates of height, based on Figure 15.

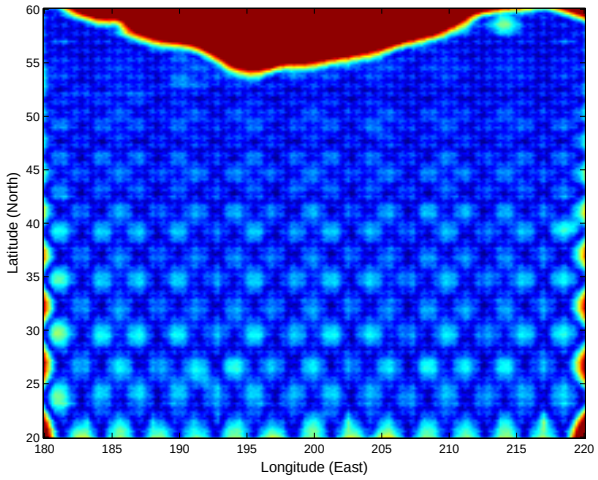

Figure 17: Error statistics corresponding to Figure 16.

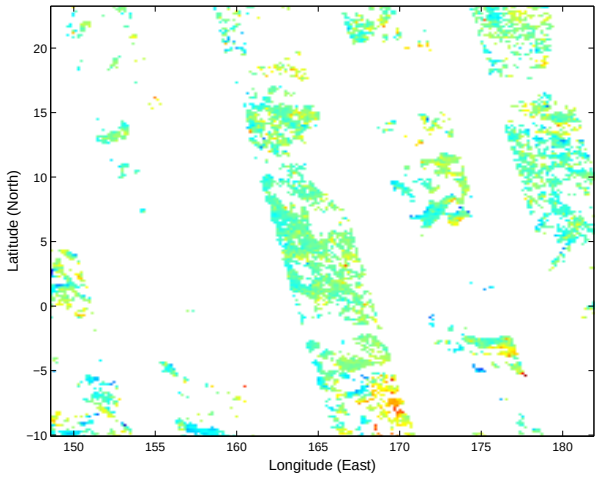

Figure 18: Surface-temperature measurements on the equatorial Pacifi c.

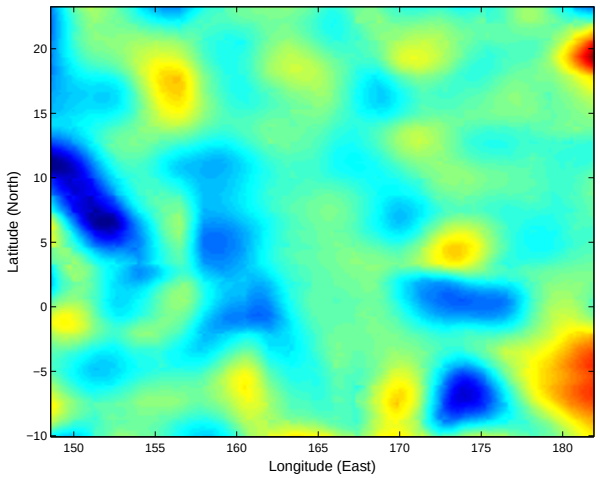

Figure 19: Temperature estimates based on a spatial Gaussian correlation prior, correlation length of 5 degrees, and the measurements of Figure 18.

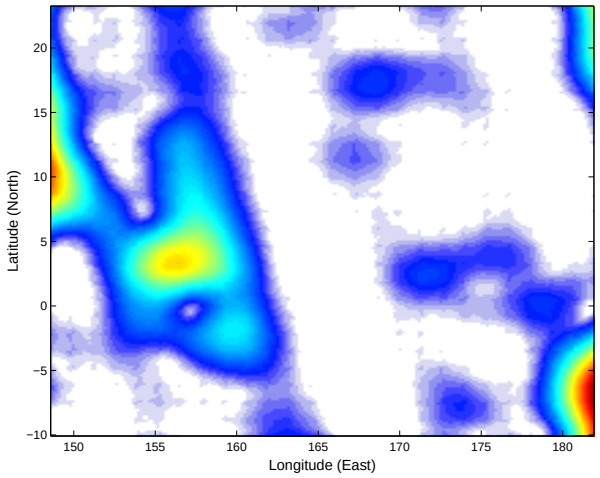

Figure 20: Estimation error statistics corresponding to Figure 19.

List of Table Captions:

1. Matrix condition number of the state having the largest dimension for a two-dimensional process with spatial Gaussian statistics, correlation length of 60 pixels. The condition number is plotted as a function of problem size ($n = 2^{\mathrm{Scales}-1}$) and state density ζ; blank entries are numerically unstable. The bottom row shows the corresponding values for the proposed method, (nearly) independent of problem size.

2. Model statistical degradation (desired correlation / realized correlation) for two pixels straddling a coarse tree boundary (as in Figure 2). The degradation is plotted as a function of problem size ($n = 2^{\mathrm{Scales}-1}$) and state density ζ; blank entries are numerically unstable. The bottom row shows the corresponding values for the proposed method, independent of size.

3. Reduction of computational effort of the proposed approach over the standard multiscale algorithm (measured via raw FLOP count). As expected, the benefit increases for more difficult problems: the reduction factor increases for larger trees and more finely sampled domains. The results are based on the tree-like prior of Figure 9.

List of Figure Captions:

1. A dense set of boundary pixels, which would conditionally decorrelate the four quadrants of a first-order Markov random field.

2. Two nodes, s_1 and s_2, neighbors in physical space, but distantly separated in tree space.

3. One possible reduced-order approximation to the state of Figure 1.

4. For how large a domain is divide-and-conquer computationally and numerically feasible . . . ?

5. Illustration of the first three levels of a quad-tree.

6. The estimation of a Markov random field using a multiscale tree based on a reduced-order model, as in Figure 3. Note the appearance of estimation artifacts at several tree boundaries.

7. An alternative choice of reduced state, appropriate for estimating the upper-left quadrant of a process.

8. Estimates of a Markov random field, based on a single-quadrant model (such as in Figure 7).

9. The mosaic of four sets of estimates, one set of estimates per quadrant as shown in Figure 8. Note the absence of artifacts.

10. Proposed approach: construct a hierarchy of statistical representations, such that we estimate $f \times f$ region **A** based on a detailed representation of **B** and a more moderate one of **C**; g, controlling the width of **B**, is related to the correlation length of the process.

11. Specific approach: the problem is broken into $2^{d(\bar{m})}$ pieces on scale $\bar{m}$, where $\bar{m}$ is selected to keep the size $n2^{-\bar{m}}$ of each piece constant, independent of total problem size n. Each piece **A** is surrounded by two concentric bands of decreasing statistical detail ζ_A, ζ_B, followed by a coarse representation **C** of the remaining domain.

12. Example multiscale implementation for a 1D example; below the fourth scale the tree is regular dyadic.

13. Example multiscale implementation for a 2D example; the structure is a regular quadtree below the fourth scale. The number within each node indicates the number of descendents. Note that this type of model is highly redundant; for example, the state at node α will be used in nine different models.

14. The merging of three trees; all arcs are followed for the upwards pass, only the thick arcs are followed downwards.

15. Measurements of ocean height in the northern Pacific.

16. Dense estimates of height, based on Figure 15.

17. Error statistics corresponding to Figure 16.

18. Surface-temperature measurements on the equatorial Pacific.

19. Temperature estimates based on a spatial Gaussian correlation prior, correlation length of 5 degrees, and the measurements of Figure 18.

20. Estimation error statistics corresponding to Figure 19.