Ecological Applications, 21(5), 2011, pp. 1523–1536

Inferential ecosystem models, from network data to prediction

James S. Clark,[1,2,6] Pankaj Agarwal,[3] David M. Bell,[1] Paul G. Flikkema,[4] Alan Gelfand,[5] Xuanlong Nguyen,[5] Eric Ward,[1] and Jun Yang[3]

[1]*Nicholas School of the Environment, Duke University, Durham, North Carolina 27708 USA*
[2]*Department of Biology, Duke University, Durham, North Carolina 27708 USA*
[3]*Department of Computer Science, Duke University, Durham, North Carolina 27708 USA*
[4]*Department of Electrical Engineering and Computer Science, Northern Arizona University, Flagstaff, Arizona 86011 USA*
[5]*Department of Statistical Science, Duke University, Durham, North Carolina 27708 USA*

Abstract. Recent developments suggest that predictive modeling could begin to play a larger role not only for data analysis, but also for data collection. We address the example of efficient wireless sensor networks, where inferential ecosystem models can be used to weigh the value of an observation against the cost of data collection. Transmission costs make observations "expensive"; networks will typically be deployed in remote locations without access to infrastructure (e.g., power). The capacity to sample intensively makes sensor networks valuable, but high-frequency data are informative only at specific times and locations. Sampling intervals will range from meters and seconds to landscapes and years, depending on the process, the current states of the system, the uncertainty about those states, and the perceived potential for rapid change. Given that intensive sampling is sometimes critical, but more often wasteful, how do we develop tools to control the measurement and transmission processes?

We address the potential of data collection controlled and/or supplemented by inferential ecosystem models. In a given model, the value of an observation can be evaluated in terms of its contribution to estimates of state variables and important parameters. There will be more than one model applied to network data that will include as state variables water, carbon, energy balance, biogeochemistry, tree ecophysiology, and forest demographic processes. The value of an observation will depend on the application. Inference is needed to weigh the contributions against transmission cost. Network control must be dynamic and driven by models capable of learning about both the environment and the network. We discuss application of Bayesian inference to model data from a developing sensor network as a basis for controlling the measurement and transmission processes. Our examples involve soil moisture and sap flux, but we discuss broader application of the approach, including its implications for network design.

Key words: Bayesian prediction; carbon–energy–water balance; ecosystem data; research networks; sensor networks.

Introduction

Forecasts based on models informed by data and scenarios for change are a goal of ecosystem science. As an example, predicted water use by plants in a more arid climate might require observations obtained from a range of climates with anticipated boundary conditions, including soil properties, future water supply, and atmospheric demand. Predictions could be inaccurate for many reasons. Extrapolation beyond the observed boundary conditions might be unavoidable: the range of climates studied today might not include the combinations of soils and climates that will prevail in the future. Critical processes might not be observed at the appropriate scale, prompting, for example, use of what is known about leaf response to weather as a surrogate for canopy response to climate. Finally, it may be hard to integrate observations in a way that allows for coherent probabilistic statements about predictions.

Prediction is not solely about observations in the future; the same issues and techniques can apply to observations from the past or those that could be obtained now. Indeed, accurate predictions could allow informed decisions about how to collect data efficiently. Here we consider prediction concepts that could help meet the challenges and opportunities of network data. We point out that prediction concepts apply at all stages, from data collection to ecological forecasts, regardless of whether or not predictions apply to the

Manuscript received 9 July 2009; revised 3 June 2010; accepted 7 June 2010; final version received 12 August 2010. Corresponding Editor: D. S. Schimel. For reprints of this Invited Feature, see footnote 1, p. 1427.
[6] E-mail: jimclark@duke.edu

future. We use simulated and actual examples to show how prediction could improve the efficiency of data collection and coherency for ecological forecasts.

The current and future expansion of networked data acquisition motivates us to focus on specific challenges and possibilities of sensor networks (Martinez et al. 2004, Pottie and Kaiser 2005), but the issues are general. The potential of sensor network data is recognized by agencies that consider their role in national research platforms NEON and CLEANER (Collins et al. 2006). Network data sufficiently dense in space and time could address such problems as behavioral responses to weather and climate, population dynamic consequences of environmental heterogeneity, and element and energy transfers between the atmosphere and biosphere (Martinez et al. 2004, Szewczyk et al. 2004, Porter et al. 2005). With this promise comes challenges. As with remote sensing, weather, and climate, the size of data sets can be daunting. Data streams from sensor networks can be difficult to transmit, store, mine, analyze, and comprehend (Lane 1997, Pottie 2001). As a simple example, understory light levels monitored by sensor networks could provide an unprecedented perspective on photosynthetic responses to sun flecks at scales of minutes (e.g., Naumburg et al. 2001) and climate change impacts on phenology across years (White and Nemani 2006, Zhang et al. 2006). To satisfy both goals implies sub-minute-scale observations spanning up to a decade or more, at 1 051 200 observations per sensor per year. Batalin et al. (2005) point out that adequate spatiotemporal coverage of sunflecks could involve 10^4 samples for an area as small as 1000 m^2. Because these are "raw data," clusters of observations would have associated metadata that would need to be consulted prior to analysis, potentially including sensor performance, battery life, missed transmissions, and so on. Although climate–phenology models might get by with less frequent observations, and sunfleck-photosynthetic models might need high-frequency data only now and then (e.g., certainly not at night), satisfying both research needs requires massive data storage and sophisticated data mining.

Second, data collected at times and places deemed most "informative" might not match those desired by a user who would like to assimilate them into predictive models: ecosystem models often require data distributed uniformly in space and time. However, even coverage in geographic space and through time might not provide adequate coverage of covariate space. Future combinations of soils and climates might not correspond to those observed today. At a different scale, soil moisture is more often near field capacity and wilting point than anywhere in between, yet field capacity and wilting point are "most predictable" and require the lowest sampling density.

We consider how the emergence of networked data collection, particularly with wireless networks, might benefit from a flexible concept of data and models, motivated by the need to eventually use those data for prediction. This goal represents a direct application of the forecasting techniques that are the focus of this Invited Feature. We begin by suggesting that ecologists may not need or even want observations as dense as traditionally thought. In wireless networks, dense data collection means not only redundancy, but also rapid battery depletion, decreased sensor life, and labor to maintain networks (Cardell-Oliver et al. 2005). These costs of data collection are specific to wireless networks, but all data collection efforts entail cost–benefit tradeoffs. Obviously, there is need for capacity to sample densely (e.g., rapid response to sunflecks), but it would be valuable to do so only when such measurements are informative. We then emphasize that "raw" data may not be the most important product from sensor networks. We note that many environmental "data" are already highly processed before they are made available to users. Acceptance of (1) the differential value of observations and (2) modeling products as data leads inevitably to the notion of inferential ecosystem models as arbiters of what is worth collecting. The value of an observation depends on what can be learned from it, which depends on a model. The idea of using predictive models for data collection, rather than solely for processing data already in hand, shifts attention to how to control data collection in such a way that it does not preclude eventual use of those data in models that will be developed for many purposes. Reactive algorithms that initiate dense sampling following events (e.g., rainfall Batalin et al. 2005, Cardell-Oliver et al. 2005) provide partial solutions.

Our goal of assessing the value of an observation goes beyond identifying when a variable is changing to focus instead on the learning that derives from it. It relies on capacity to "predict" observations, as opposed to, say, estimate parameters in a specific model: observations are the important product for a broad community, whereas parameter estimates are model (and thus user) dependent. Data suppression is then based on suppressing the predictable, knowing that predictable observations can be reconstructed. Models can be simple enough to benefit from the important relationships that facilitate prediction, while having few estimated parameters. The products of the network can be predictive means and variances, which accommodate not only measurement error, but also the unknowns associated with processing. These predictive means and variances can be uniform in space and time, despite high selectivity in raw data collection.

The ecosystem models we use as examples differ from those common in the literature in that they are simple enough to be transparent and provide predictions that could be used in many models, and they are constructed with coherency in mind. Here the term coherence refers to the fact that there is a joint posterior distribution of all parameters and latent states in the model. The uncertainty in model outputs is conditional on the

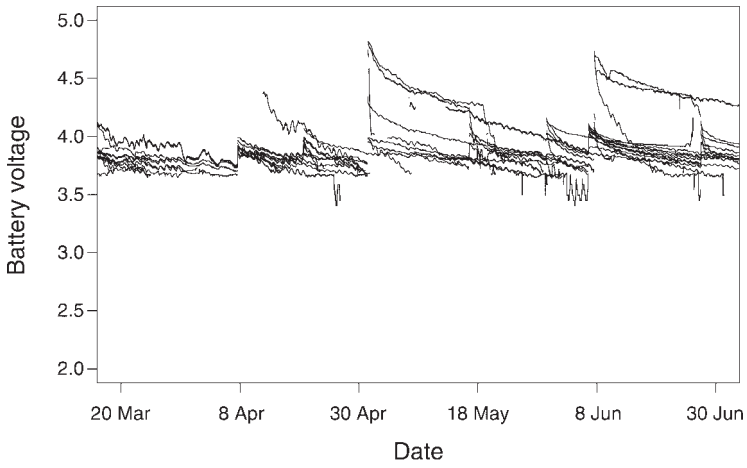

FIG. 1. Battery life plot shows voltage depletion and replacement (step increases in voltage) every few weeks in 2008. Different rates of depletion depend on network traffic at each node.

uncertainty of model inputs and the model itself, and it can be expressed as a credible interval with probabilistic interpretation. Analysis involves only widely accepted distribution theory. The presented models are intended only as examples to highlight key concepts.

We begin by summarizing challenges of network data collection from the perspective of its eventual use in predictive models. We then use simulated examples to show how inferential ecosystem models can make data collection efficient and their role in predicting not only state variables of interest but also the missing data that might be needed by a range of different investigators. An application to transpiration shows the capacity to transform incomplete information to data products that are amenable to analysis by a broad user community.

Challenges of Dense Network Data

Dense data streams are expensive to obtain, to extract, and to interpret, the latter because data sets can be uneven, interrupted, and/or gap filled with values lacking estimates of uncertainty. Here we summarize costs and benefits of wireless sensor network data, emphasizing their relationship to what might be actually used by ecologists, involving predictive modeling.

In wireless networks, costs of frequent data transmission include battery depletion, network congestion, data losses, and maintenance costs that come with reduced sensor life (Ganesan et al. 2004). As one example, battery lifetimes in previous deployments of the WiSARDNet (Flikkema et al. 2006) range from several weeks to months (Fig. 1). As with any data collection effort, there is a tradeoff involving the potential value of high-density observations vs. the benefits that might come from additional effort devoted elsewhere. Given that wireless transmissions often dominate the energy cost of sensor node operation, censoring of transmissions can dramatically lengthen battery lifetimes. Hence the tradeoff between dense measurements vs. battery longevity and maintenance of the network motivates exploration of the value of an observation: one that weighs value against cost. Here are a few prediction concepts that can help address this tradeoff.

Consider a stream of data $\{y_i\} = \{y_{i,1}, y_{i,2}, \ldots, y_{i,t}\}$ at location i at times $t = 1, 2, \ldots$. The value of an observation could be gauged by the extent to which it reduces the uncertainty about a parameter we wish to estimate or about a prediction of $y_{i,t}$ at some different time t or some location i' where data were not collected. A posterior distribution of a parameter, call it θ, which comes from a model for data $\{y\}$, is

$$p(\theta \,|\, \{y\}, m) \propto p(\{y\} \,|\, \theta, m) p(\theta). \quad (1a)$$

On the right-hand side are the likelihood and prior. Both inference and prediction depend on one or more models m. We consider explicit examples later. For now consider that prediction is a natural extension, involving an integral:

$$p(\{y'\} \,|\, m, \{y\}) = \int p(\{y'\} \,|\, \theta, m) p(\theta \,|\, \{y\}, m)\, d\theta \quad (1b)$$

where the integrand is the likelihood structure and posterior, respectively. We arrive at a predictive distribution of $\{y'\}$ on the left-hand side, which is conditioned on the model and on the data $\{y\}$. This predictive framework is widely used, to predict the data themselves (e.g., cross validation, predictive loss), to transform unevenly spaced observations to uniform grids (e.g., kriging), and to predict values at different times (e.g., hindcasting or forecasting). The questions concerning the value of an observation can now be framed as (1) how much would an additional observation affect our estimate of θ and/or (2) our ability to predict $\{y'\}$? Before discussing an application, we consider the role of the data $\{y\}$ and the model(s) m.

The conundrum of dense data of limited value

The amount of information contained in an observation is limited if it is redundant (i.e., if it could have been predicted) based on what is known of the process from observations and prior analysis. Four of the state variables monitored in the Duke Forest WISARDNet (Fig. 2) show different levels of predictability. Soil moisture is a "slow" variable, changing rapidly during and shortly after rainfall events and more slowly thereafter. Between rainfall events, change in soil moisture can be predictable not only because it changes slowly, but also because its behavior is strongly influenced by temperature and soil properties that govern how fast it can change. The nearly flat portions of curves in Fig. 2a result from the fact that soil moisture is increasingly difficult to extract as it approaches the "wilting point." Clearly knowledge of when it rains could help predict soil moisture loss. Predictability of soil moisture loss at a location typically contrasts with high spatial variation.

Light availability in the forest understory (Fig. 2b) is an example of a variable that changes rapidly and thus might be collected at dense intervals. Short-term sunflecks can have a large impact on carbon assimilation by plants, depending on the time it takes to activate the photosynthetic apparatus: the larger the sunfleck, the longer the duration and the larger the effect. Ecologists devote substantial effort to measure rapid fluctuations with minute-scale observations (Naumburg et al. 2001, Collins et al. 2006). Yet for larger sunflecks even this "fast" variable is predictable based on some simple relationships; accurate predictions require knowledge of light availability above the canopy at the time in question and at the same location on previous days. The predictability comes from the fact that shadows fall in about the same location day after day (Fig. 3). Their locations change slowly during the growing season, and that too is predictable from solar geometry. Ephemeral clouds can be monitored by a single sensor open to full sun. As soon as the relationship is established between light above the canopy and at a sensor location, cloud effects observed with an above-canopy sensor become predictable, using knowledge gained from transmission obtained on previous days. Granted, coherent representation of the uncertainty can be challenging, but all of the components of Eq. 1 are available.

Precipitation is comparatively erratic and essentially unpredictable based on the recent past. However, knowledge of precipitation at one location (say, above the canopy; Fig. 2c) provides information about soil moisture change at other locations (e.g., Fig. 2a). Relative humidity (Fig. 2d) is somewhat predictable, with a clear diel cycle, but varies substantially from one day to the next. In short, temporal variation in some meteorological variables (e.g., VPD), which can be monitored at one landscape location, can far exceed their spatial variability across nodes within a data network. Taken together, variables exhibit variation on a range of scales, and they differ in terms of predictability. The more predictable the variable, the less is gained from an observation, in the sense that it does not contribute information that is not already in hand.

In addition to data, the predictive distribution requires one or more models. Before discussing a strategy for inferential modeling to guide data collection, we first recognize that models are already an important part of ecological data collection: it is not the use of models that is new here, but rather their use in a predictive mode.

"Raw" data are already cooked

There is a long tradition in the environmental sciences of confidence in data and skepticism of models. We argue that a program involving predictive data collection can include a sparse set of observations supplemented by a dense predictive distribution. In other words, the densely spaced data needed for ecosystem models would come from models that assimilate data. This would be a departure from current practice, which focuses on raw data as inputs. Motivation comes in part from the challenges of collection, storage, and retrieval of massive data sets.

Large-scale efforts could entail massive quantities of raw data, much of which may be difficult to extract, and it will include unknown errors. Mining the data for domains of interest will be difficult. It could be overwhelming to mine even the metadata for clues on uncertainty that should impact analysis. We begin by pointing out that (1) many of the "raw" data accumulated by ecologists are already the products of models and (2) analysis of model products is often not fundamentally different from analysis of data.

Most observations obtained with the aid of an instrument involve both measurement and processing. Transducers for most environmental variables quantify voltage or current, which are translated by a model. The models range from empirical calibrations to physical models. The transducers themselves are noisy, can drift over time, and are prone to numerous failure modes. For example, time domain reflectometry (TDR) does not directly measure soil moisture but rather the travel time of an electromagnetic wave traversing the soil medium over the length of the TDR rod. Water molecules contribute most to the dielectric properties and hence are mainly responsible for slowing the electromagnetic wave. Empirical calibrations are used to relate dielectric properties of the medium and soil moisture, often expressed as a polynomial. Errors include drift from these idealized curves when factors such as salinity or charged clay particles modulate propagation of the electromagnetic wave. Likewise, remote sensing products are highly processed, both spectrally and spatially. GIS layers often derive from spatial models and classification algorithms, and they involve transformations to rectify remotely sensed images.

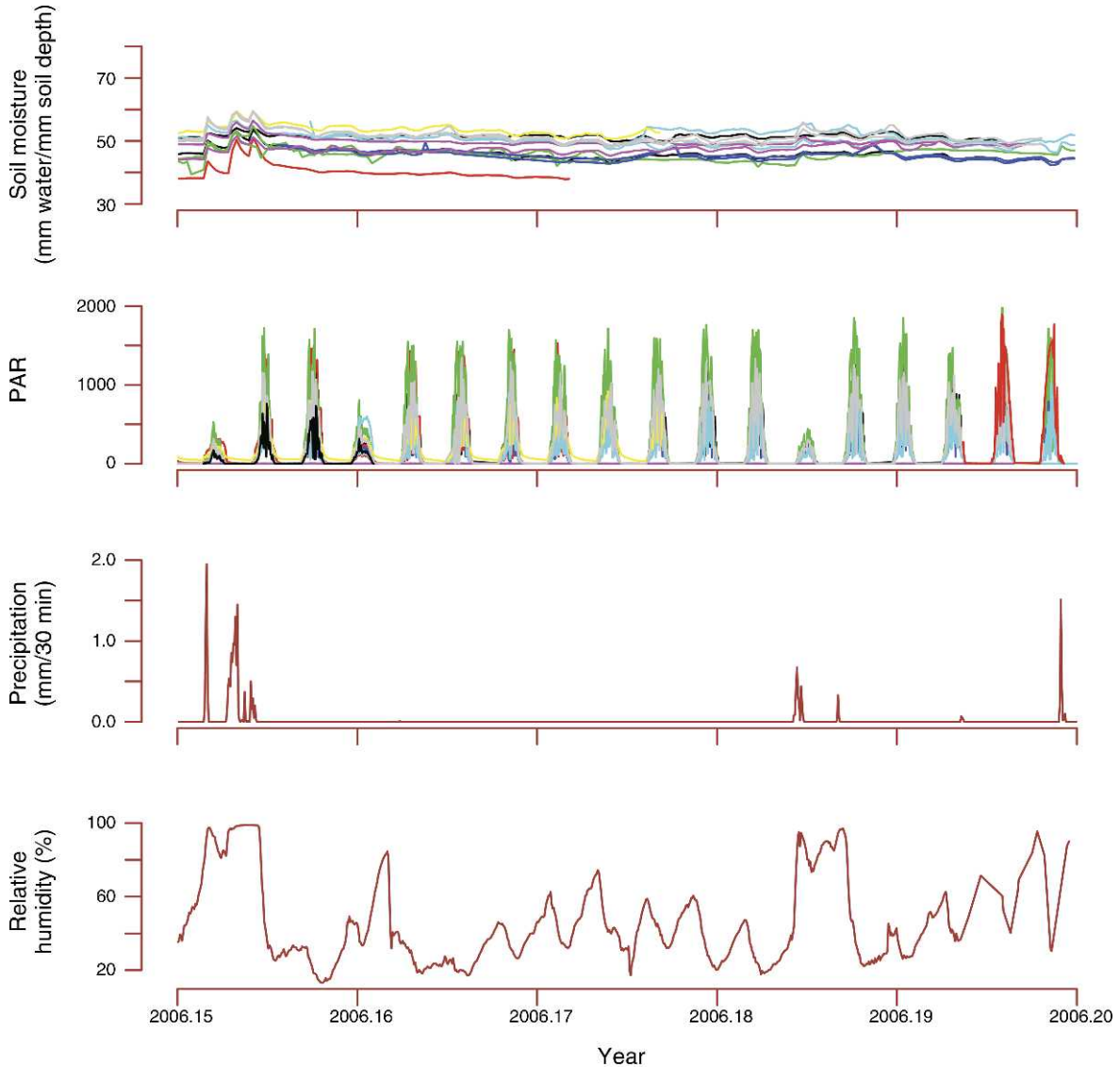

FIG. 2. Examples of four variables measured in the Duke Forest wireless sensor network. For panels with multiple curves, the variable is measured at multiple nodes (soil moisture and photosynthetically active radiation [PAR]). Those with single curves are measured at a single, centrally located tower (precipitation, relative humidity). The short sequence here spans 18 days. Soil moisture was measured using Decagon EcH2Oprobe EC-20 capacitance-based transducers (Pullman, Washington, USA); photosynthetically active radiation was sensed using GaAsP photodiodes (Hamamatsu, Japan); and a Vaisala WXT510 integrated weather station (Helsinki, Finland) sampled precipitation and relative humidity using piezoelectric and capacitive sensors, respectively. Numbers on the *x*-axis represent fractions of the year 2006.

Post-processing is no longer just an option for many applications, often being required to correct or even generate data that were missed. As an example of the latter, gap-filling procedures are routinely applied to generate continuous sequences in eddy-covariance measurements. Models are calibrated for conditions when eddy-covariance measurements are known to approximately document ecosystem sources and sinks, and are used to predict missing data when eddy-covariance measurements are absent (due to sensor failure) or their basic assumptions are violated (flow is not fully developed turbulence, advective fluxes do not balance the depth-integrated sources from fluxes, or the high-frequency spectra losses are large; see Cava et al. 2004, Richardson et al. 2006, Stoy et al. 2006).

Faulty or anomalous data (Ni et al. 2009) can arise from transducer failures (e.g., short or open circuits caused by weather, aging, or damage by animals). These are often expressed as stuck/constant data streams. Transducer hardware may drift due to aging, introducing bias, or have soft failure modes that result in noisy or wandering readings. Over time, they are replaced with

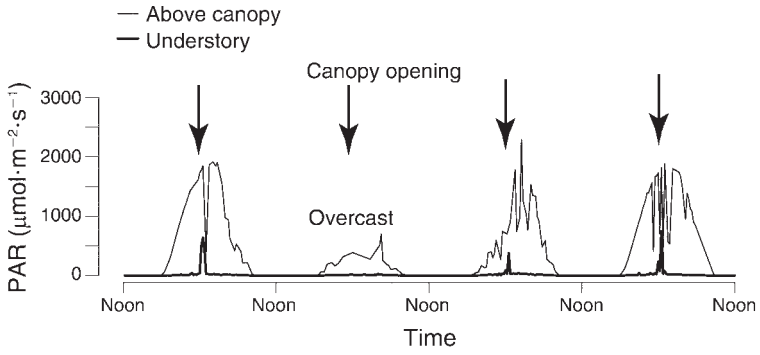

FIG. 3. Light availability in the understory can be predictable, because canopy openings cast shadows at similar times each day, progressing with the season.

new technologies having different properties. Data may also be missing due to depletion of batteries, a hardware/software fault in the sensor node, or human error.

Given that many environmental data already involve integrating observations and models, it should not be controversial to suggest that future, massive data collection efforts could not only embrace, but also formalize, inferential modeling as a basis for *data collection*, not just analysis. Wireless networks should particularly benefit from this effort. Statistical models assume that data include a random element, acknowledging that observations differ from the underlying "truth" by uncertainty in measurement. Modeling can account for these uncertainties. In our examples, we emphasize control of costly wireless transmissions, estimation of missing data, and drift of measurements; however, in our experience, other types of failures (e.g., transducer faults) can be easily detected or modeled similarly.

In summary, raw data, processed data, and models share systematic bias, uncertainty, and missingness. Raw data are not necessarily more accurate, more precise, or more useful than processed versions of those data, which can combine external knowledge of interactions and measurements and yield estimates of uncertainty. Recognition of these relations suggests not only a role for interference and ecosystem models at the data collection stage, but also ways to exploit inferential tools to make data collection efficient.

Bayesian Learning for Data Collection and Prediction

Bayesian learning entails updating knowledge as data accumulate. Although results are specific to the model that is applied (Eq. 1), sensitivity to a specific model should be minimized. Models change, tending to become more complex as understanding improves and accumulating observations allow for inference on a broader range of state variables and parameters. For purposes of data prediction we desire a minimal model aimed at exploiting relationships that aid prediction of state variables. In a soil moisture example that follows, the model is limited to simple relationships involving atmospheric demand and current soil moisture, relative to field capacity, wilting point, and several fitted parameters. Although a water balance model could involve many more parameters, predicting soil moisture does not require a highly complex model.

Models help determine the value of an observation in terms of its contribution to prediction: if an observation adheres to a pattern that can be predicted based on previous observations and the model, then it has limited value. We could use the width of the predictive interval, say a quantile for the predictive density in Eq. 1b, as a criterion. We are interested not only in how well we could have predicted the observation, but also in how much it contributes to the prediction of others. This contribution comes indirectly, through its contribution to estimates of parameters in the model. The sensitivity of the posterior (Eq. 1a) to an observation, in part, determines how the observation contributes to prediction of other observations (Eq. 1b). Once parameter values are well described, in terms of a narrow posterior, additional observations may not have much additional impact on prediction of new observations. Thus, predictive distributions of data become an important element of our approach. The approach assumes in-network capacity for minimal computation (Collins et al. 2006, Flikkema et al. 2006, McIntire et al. 2006) to implement the prediction/decision algorithms that control data transmission. Because the approach we describe is based on computation that occurs primarily outside the network (parameter estimation), there are no new hardware requirements. Our approach uses plug-in values for a simple model and decisions based on thresholding of prediction error. Algorithms are easily implemented in the low-power processor chips typically used in wireless environmental sensor nodes. The energy saved can be substantial: the required processing uses roughly 2% of the energy that is required for communication of the data between two sensor nodes. Because the data must be conveyed from sensor node to node on the way to the destination, with each hop using

additional energy, the savings can quickly increase with the size of the network.

Controlling the process by suppressing the predictable

Here we provide an example of how inferential ecosystem models can control efficient data collection, while simultaneously assuring that underlying variables of interest are represented to a desired precision. There is a huge literature on data suppression and compression, based on distributed statistical signal processing and information theory (Xiao et al. 2006). Our goal is to modify ecosystem models similar to those that will ingest data as basis for data collection. Of course, the full range of models to which sensor network data will be applied in the future is unknown at the time of data collection, a point to which we return below. For now we simply state that the details of a model used at the time of data collection can impose few constraints on the future value of data, largely because models used for data collection are principally concerned with data prediction (e.g., soil moisture), rather than the parameters for a specific model (e.g., rate parameters).

As a specific example, let y_t represent soil moisture at time t that increases when there is precipitation p_t and declines due to evapotranspiration demand E and drainage D. The model is intentionally simple, describing a one-dimensional integrated soil column. Our goal is to track soil moisture (within the root zone) over time, using both data and a model to help identify "valuable" observations, ones that contribute important information. By learning about both the model and the sensors, we will determine how to transmit sparse data while maintaining acceptable (and known) uncertainty. The sparse data stream will prolong battery life, while allowing us to "reconstruct" soil moisture at dense intervals. A simple rule applied here is to transmit only the unpredictable observations.

A simple dynamic model for soil moisture is

$$y_{t+1} = f(y_t, p_t; \theta)e^{v_t} \tag{2a}$$

$$f(y_t, p_t; \theta) = y_t + p_t - E(y_t; \theta_1, \theta_2) - D(y_t; \theta_3) \tag{2b}$$

$$v_t \sim \mathcal{N}(0, \sigma_\varepsilon^2) \tag{2c}$$

where the function f in Eq. 2a is the process model Eq. 2b and has terms for the previous soil moisture value, precipitation gains, and evapotranspiration and drainage losses. The exponential in Eq. 2a makes the process error lognormal. The loss terms in Eq. 2b have simple functional forms that rely on unknown parameter values for the point of incipient plant stress due to soil moisture limitations θ_1, wilting point θ_2, and field capacity θ_3. Because soil moisture models can be arbitrarily complex (future users of the data will not all apply the same models to these data), we do not focus on the interpretation of these parameter values. Their role here is to simply improve our ability to predict soil moisture by calibrating a few key relationships. In other words, wilting point and field capacity are parameters that help us predict how rapidly moisture is lost from the soil. This "process model" in Eq. 2b has error described by v_t (Eq. 2c), which quantifies its uncertainty.

In addition to the process, we require models for data collection. We have multiple sensors j, each having error described by variance σ_z^2 and drift parameter δ_j, which accumulates with time t_j since the jth sensor was calibrated. The jth sensor measures a value $z_{j,t}$. There are calibration data w_t, which we assume to be sparse and to have known error variance, but no systematic bias. In other words, the calibration data are taken to be the standard. Here are data models for the sensor observations $z_{j,t}$ and the calibration data w_t:

$$\ln(z_{j,t}) \sim \mathcal{N}\Big(\ln(y_t)\Big(1 + \delta_j(t - t_j)\Big), \sigma_z^2\Big) \tag{3a}$$

$$\delta_j \sim \mathcal{N}(0, v_\delta) \tag{3b}$$

$$v_\delta \sim \text{IG}(s_1, s_2) \tag{3c}$$

$$\ln(w_t) \sim \mathcal{N}\Big(\ln(y_t), \sigma_w^2\Big) \tag{3d}$$

where IG stands for inverse gamma density. The random specification for drift parameters (Eqs. 3b, c) means that we view them as exchangeable, being drawn at random from a population of sensors that can drift at different rates. We assume that the error variance σ_w^2 (Eq. 3d) is known, through previous calibration. The parameters of the model are estimated through calibration. This is known as a state-space model, which is hierarchical, having error in the underlying process (Eq. 2) and in the data-generating mechanism (Eq. 3).

A decision about whether to transmit an observation depends on whether it could have been predicted, within an acceptable error limit. Once we have learned about parameters in the process model 2 and in data collection 3, we predict what a sensor will measure based on past observations. Here is a simple plug-in version of the model, with hats indicating values that are represented by estimates or predictions:

$$\ln(\hat{z}_{j,t}) = \ln(\hat{y}_t)\Big(1 + \hat{\delta}_j(t - t_j)\Big) \tag{4a}$$

$$\ln(\hat{y}_t) = f(\hat{y}_{t-1}, p_{t-1}; \hat{\theta}). \tag{4b}$$

In Eq. 4a, the soil moisture y_t is predicted based on the deterministic part of data model (Eq. 3a), the previous prediction for y_{t-1} (Eq. 2b), and the best available estimates of parameters. Thus, the predicted measurement $\hat{z}_{j,t}$ relies on the current estimate of the sensor's accumulated drift. Now if the measurement $z_{j,t}$ falls outside the acceptable bound for error,

$$|\hat{z}_{j,t} - z_{j,t}| > \varepsilon$$

then the observation is transmitted. If not, it is

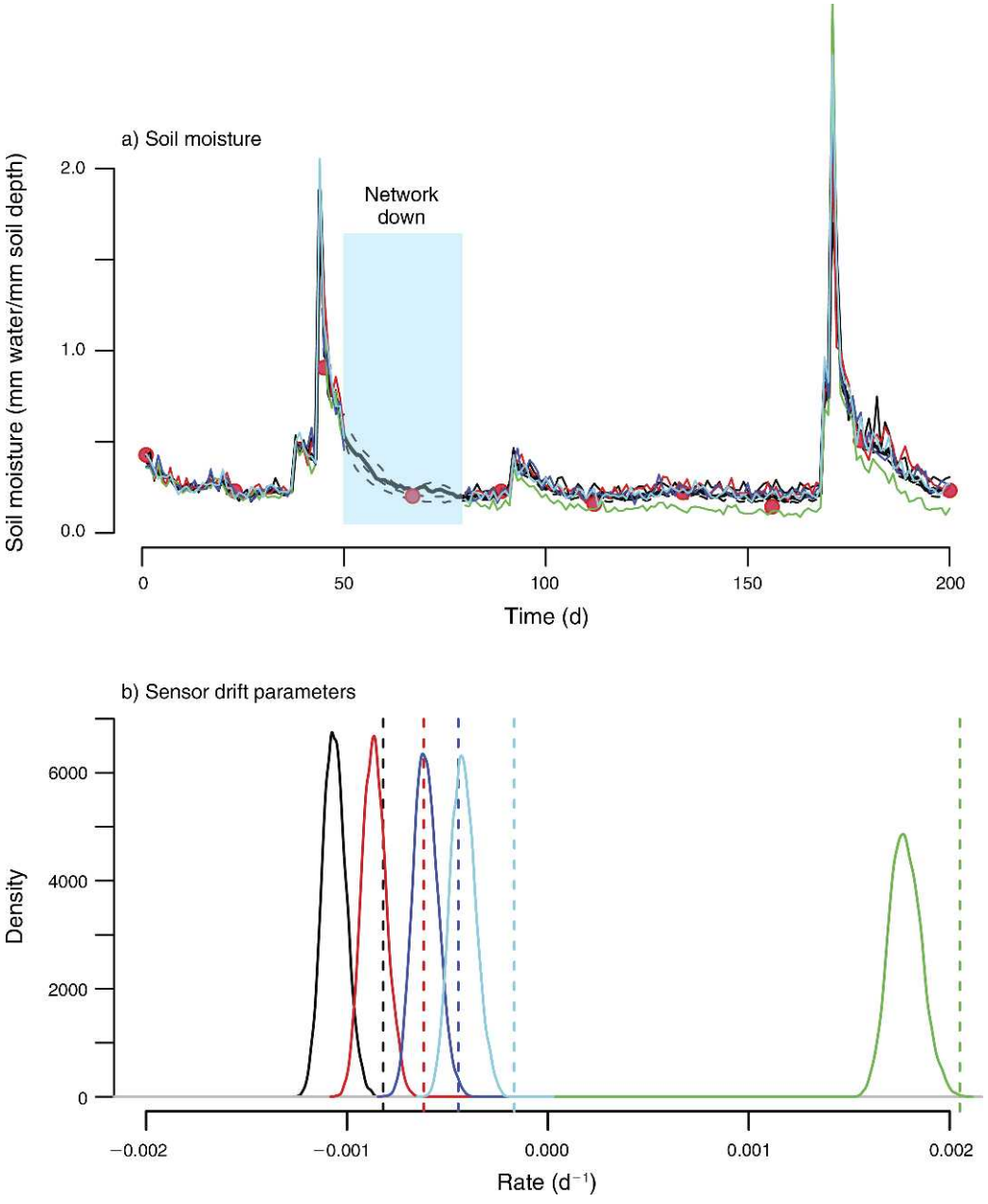

FIG. 4. (a) Simulated data using the model from Eq. 2. The underlying (unknown) soil moisture is shown as a solid black line. Each of five sensors is shown as a solid, colored line. The lines diverge from the true values over time, due to sensor drift. The calibration data are intermittent and are shown as red circles. No data are available for the period shaded in blue. The black dashed lines are 90% predictive intervals from the hierarchical Bayes model. (b) Posteriors for drift parameters corresponding to each of five sensors. Vertical dashed lines are volumes used to simulate data.

suppressed. The smaller the value of parameter ε, the more observations will be transmitted, and vice versa.

Implementation requires (1) a "training period" during which we learn about parameter values, through modeling of the data collected; (2) transmission of those parameter values to individual nodes; (3) new data transmission only for "unpredictable" observations; and (4) modeling to recover the unobserved changes in soil moisture. If too many transmissions are suppressed, then the value of ε can be reduced. Here we describe these steps in the context of a simulated process and data as described by the model above (Fig. 4a). The underlying soil moisture is shown as a solid black line. Five sensors are shown in different colors, with calibration data as red dots. To emphasize that the approach does not depend on a flawless network, we also assume that the entire network is down for a time interval.

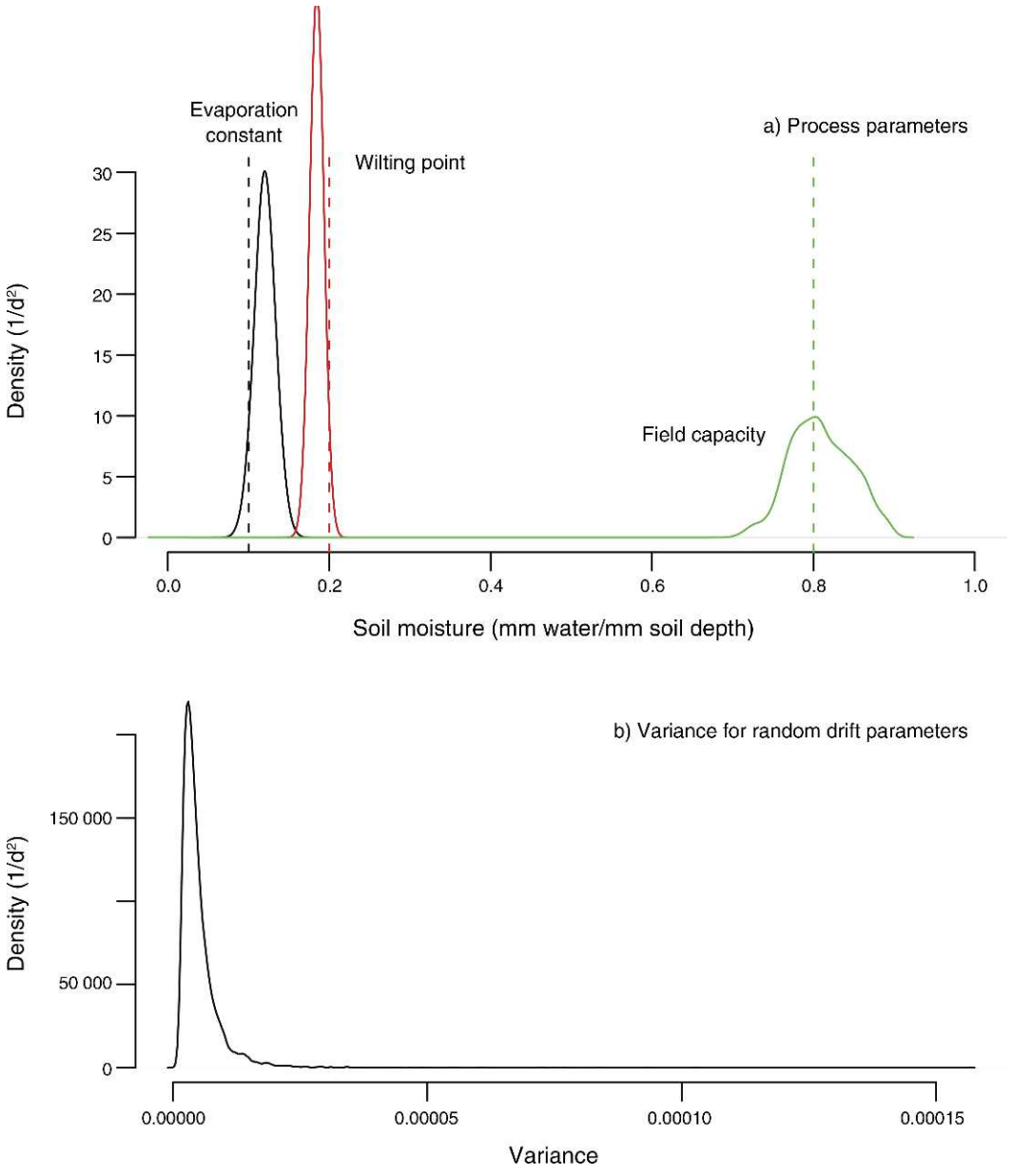

FIG. 5. (a) Posterior estimates for process parameters that control evapotranspiration rate and (b) variance for the drift parameters. These are largely nuisance parameters, needed primarily for prediction of soil moisture.

Training period.—Data collected from the network are used to estimate parameters. The full posterior is $p(\{y\}, \theta, \{\delta\}, \sigma_\varepsilon^2, \sigma_z^2 \mid \{z\}, \{w\}, \{p\}, \sigma_w^2)$, showing the variables and parameters that must be estimated to the left of the vertical bar, and those that are known to the right of the vertical bar. Marginal posteriors for the parameters in this example are shown in Figs. 4 and 5. The unknown states $\{y\}$ (Fig. 4a), process parameters θ (Fig. 5), and parameters from the data models (Fig. 4b) can be estimated from the data observed by the network. The 95% predictive intervals for soil moisture (dashed lines in Fig. 4a) show that, despite sensor drift and even complete network failure, soil moisture can be accurately predicted. For this particular example, the estimates of drift parameters are somewhat biased (Fig. 5b), but these parameters are not the main goal of the analysis. Most importantly, their specific values do not have large impact on predictive capacity (Fig. 4a).

Updating nodes.—To make predictions, sensor nodes within the network need only the capacity to plug parameter values into Eqs. 4a and b. The frequency of updating nodes with new parameter estimates can be based on experience with "acceptable error." For this

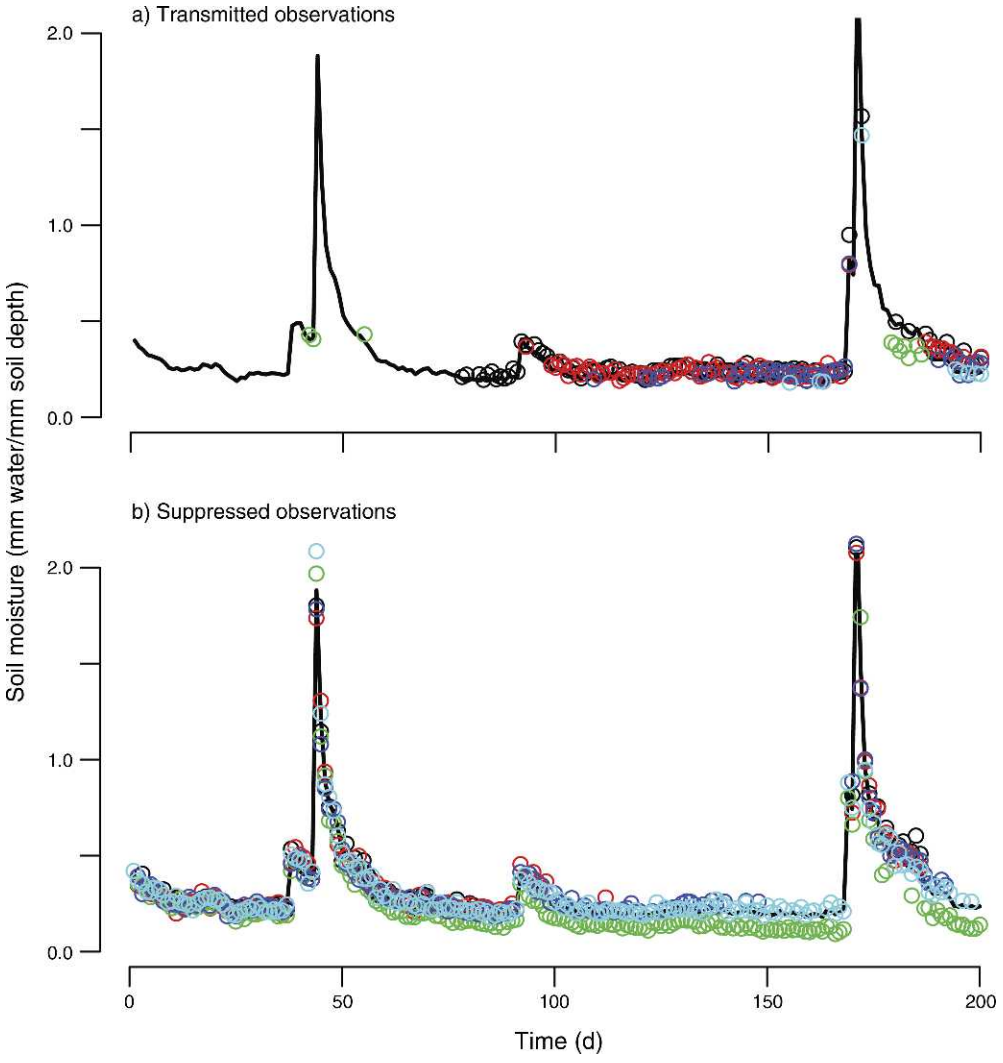

FIG. 6. (a) Observations sent (because they fell outside the prediction window) and (b) suppressed (because they could be predicted). Colors match those for sensors in Fig. 2. Increasing numbers of observations are transmitted over time as sensors drift out of calibration, a source of error that can be estimated and known at each sensor node.

example, we will see that predictive capacity declines with accumulated sensor drift.

Efficient transmission.—Based on the updated parameter values and plug-in predictions from Eq. 4, observations are transmitted or not. As emphasized above, unpredictable observations (outside the 2ε envelope) are always transmitted. Fig. 6 shows suppressed and transmitted observations for this example as dots.

Reconstructing the data.—Full posterior distributions of $\{y\}$ can be recovered by post hoc out-of-network processing, facilitated by the fact that prediction rules used to transmit data are known. Specifically, models for the missing data are constrained by knowledge now encoded by the model and data collection history, including parameter values. Consider the conditional posterior for an untransmitted sensor observation:

$$p\Big(\ln(z_{j,t})\,|\,y_t,\,\hat{z}_{j,t},\,\varepsilon,\,\ldots\Big) \propto \mathcal{N}\Big(\ln(y_t)\Big(1+\delta_j(t-t_j)\Big),\,\sigma_z^2\Big) \times I\Big((\hat{z}_{j,t}-\varepsilon)<z_{j,t}<(\hat{z}_{j,t}+\varepsilon)\Big) \quad (5)$$

where I is the indicator function, equal to 1 when its argument is true and 0 when false. This is the density we would use to describe uncertainty in the missing value $z_{j,t}$. The first density on the right hand side of Eq. 5 is potentially broad, but it is constrained by the indicator function, which relies on that fact that the suppression interval used by the sensor is known. This knowledge constrains the estimates of missing data, which, in turn, allow for more accurate prediction of soil moisture. Simulating the posterior distribution of unknowns is termed Markov chain Monte Carlo and is now in

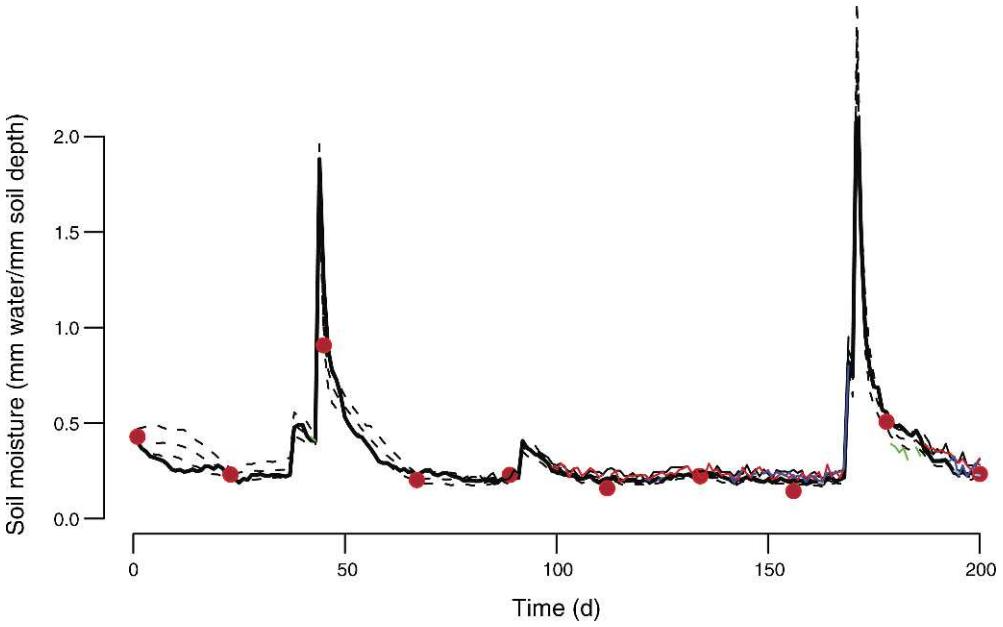

FIG. 7. Predictive intervals for soil moisture (dashed lines) based on data transmitted in Fig. 4a (and shown here as colored lines) compared with the underlying true value (solid black line) and calibration data (red dots).

common use for inference. Here we are applying it to the problem of predicting values of missing or uncertain data.

Using the suppression scheme described above results in limited transmission (Fig. 6), but does it permit us to infer the details that matter? The suppression scheme imposed here is severe, resulting in infrequent transmission despite substantial change in soil moisture level until sensor drift becomes substantial. The set of nodes that transmit most are not necessarily those with the greatest drift, but rather depends on how well the drift is estimated and, in this example, the sign of the drift. Only when accumulated drift begins to result in large errors does transmission begin to increase for all nodes (Fig. 6).

Based on sparse data we can recover detailed information on soil moisture. The data collection shown in Fig. 6 is the basis for 95% predictive intervals (dashed lines) in Fig. 7. Based on transmission of a fraction of the data we not only reconstruct the series, but can also provide uncertainties. We have intentionally ignored spatial redundancy in this example, yet there are clear opportunities for exploiting correlation to further reduce data transmissions.

Predicting transpiration

What would data provided by a scheme described here look like, and what would be the implications for forecasting water use when and where data were unavailable? D. M. Bell, E. J. Ward, R. Oren, P. G. Flikkema, and J. S. Clark (*unpublished manuscript*) and E. J. Ward, D. M. Bell, J. S. Clark, H. S. Kim, and R. Oren (*unpublished manuscript*) consider the challenge of gap-filling the discontinuous sequences of observations that are used for analysis of water balance. Transpiration rates depend on water conductance through sapwood, driven by soil moisture, temperature, light, and vapor pressure deficit (Oren and Pataki 2001). It is estimated based on sap-flux data. Granier-type transducer probes exploit the fact that the flux of water through xylem can be related to the temperature difference between a heated and an unheated probe. Sap-flux measurements at sensor nodes can be unreliable, due to battery life and several transducer failure modes. The investigator is confronted with a discontinuous set of observations that must be filled in if they are to be used in most models of water, CO_2, and energy exchange. The problem is a direct extension of the transmission suppression issues described previously: How well can missing values be predicted on the basis of a simple model and estimates of parameters for both the process and observation system?

As with soil moisture, we have a process model, in this case describing how stomatal conductance G_t is affected by vapor pressure deficit V_t, light availability Q_t, and soil moisture M_t. Stomatal conductance affects evapotransporation, which, in turn influences measured sap flux J_t (Fig. 8). Details of the model are beyond the scope of this summary and detailed in Bell et al. (*unpublished manuscript*) and in Ward et al. (*unpublished manuscript*). In brief, we apply a state-space version of the model used in (Ward et al. 2008). The counterpart for Eq. 2, i.e., the process model, could be summarized this way:

$$G_t = f(G_{t-1}, V_{t-1}, Q_{t-1}, M_{t-1}; \theta_G) + v_t$$

$$v_t \sim \mathcal{N}(0, \sigma_\varepsilon^2). \quad (6)$$

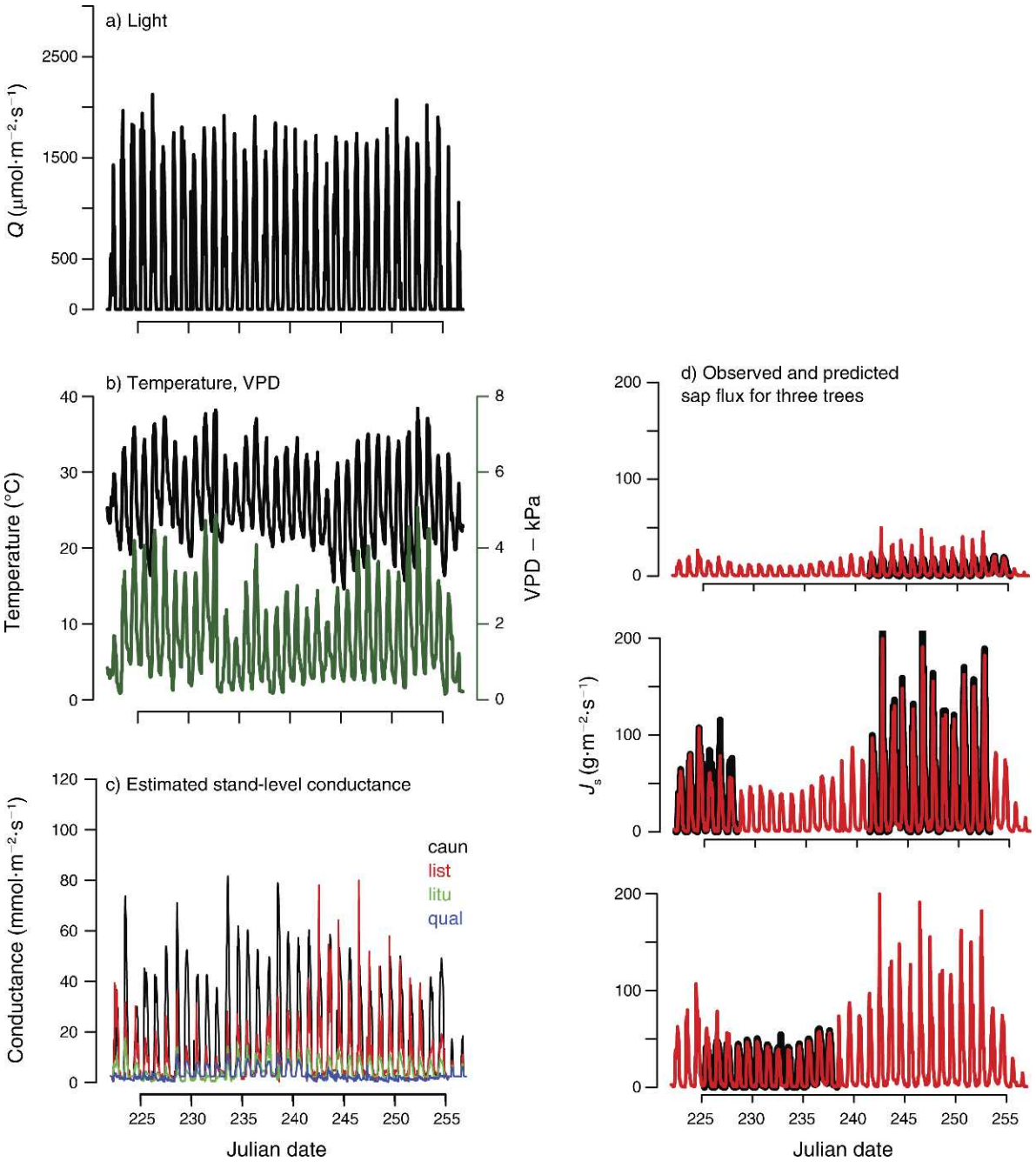

FIG. 8. (a, b) Variables measured in sensor networks, (c) estimated conductance, and (d) predicted (red) and observed (black) sap flux for three *Liquidambar* trees. Large differences between trees result in part from canopy status. Key to abbreviations: VPD, vapor pressure deficit; Q, photosynthetically available radiation; J_s, sap flux; caun, *Carya* species; list, *Liquidambar styraciflua*; litu, *Liriodendron tulipifera*; qual, *Quercus alba*. Julian date is day of year, e.g., day 1 equals 1 January.

The mean function $f(\ )$ describes how conductance is regulated by environmental variables and capacitance, i.e., the lagged influence of the previous value G_{t-1}. The counterpart for Eq. 3, the observation model for probe k in tree i, is summarized as follows:

$$J_{ik,t} \sim \mathcal{N}\left(E(G_t, V_t, T_t)\phi_{ik}s_i, \sigma_J^2\right)\frac{A_L}{A_S} \quad (7)$$

where the mean function is the product of stand evapotranspiration $E(\ )$, the effect of probe depth ϕ_{ik} and canopy stature s_i on the sap flux measurement $J_{ij,t}$. The leaf area to sapwood area ratio A_L/A_S is dimensionless.

An example application of the model fitted to 10 trees in the Duke Forest, North Carolina, is shown for three

trees in Fig. 8d. Each sequence contains a substantial gap. The underlying conductance is estimated for four different species in the stand (Fig. 8c), with the strong diel pattern responding to light (Fig. 8a), temperature (Fig. 8b), vapor pressure deficit (Fig. 8b), and soil moisture (not shown). The differences among individuals in Fig. 8c in large part reflect the different sizes of trees in the study, with large trees having more canopy exposure and, thus, more light. This is clear in Fig. 8d, where the first of three *Liquidambar* trees does not have direct exposure to sunlight. The main point for emphasis here is that strong relationships between state variables that are measured (Fig. 8a–b) and observations allow for prediction of state variables that are not observed (Fig. 8c) and data that could not be collected (Fig. 7d).

An important advantage of the modeling approach applied here is that it provides probabilistic statements not only about parameters and latent states (credible intervals are not shown in Fig. 8c to reduce clutter), but also on the observations themselves. In Fig. 9 we show the variance on predictions for conductance G_t for times when observations were obtained (black in lower panel) vs. when they were not (red). Several trees contribute information at any one time, and not surprisingly we find a decline in the variance of the posterior as numbers of trees contributing data increase (upper panel in Fig. 9).

The predictive intervals on observations are valuable, because they report uncertainty that is estimated on the basis of all data collected in the network. This information can be critical, even for informal inspection, when an investigator would like to assess the uncertainty associated with particular observations. For a formal analysis, the predictive variance can enter as a weight for the observation, determining its relative contribution.

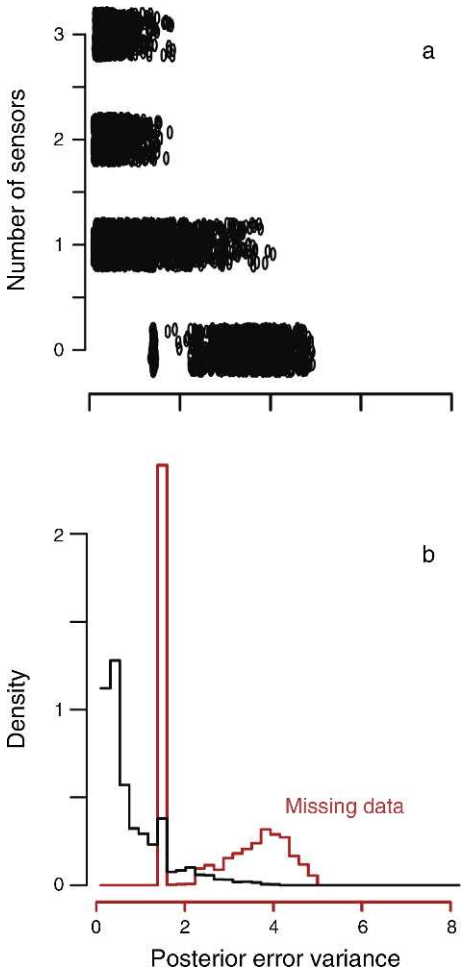

FIG. 9. Variances for posterior estimates of daytime conductance G_t plotted against (a) numbers of trees with active sensors and (b) as densities of estimates. The estimates for times of missing data (in red) are more uncertain.

EMERGING PERSPECTIVES ON DATA AND THE ROLE OF MODEL-BASED INFERENCE

The products of data collection driven by inferential models could be extended to not only (1) raw data and (2) metadata, including the model used to generate predictions, but also (3) predictive means and variances for the observations and latent states that ecologists will be most interested in using in subsequent analyses. Because models used at the data collection stage are simple, containing only relationships that help to predict missing observations, the predictive intervals are not particularly sensitive to specific model assumptions. Models can be informed by relationships that involve a small number of parameters, such as considered here, or they can rely largely on spatial covariance structure, where predictive information comes from proximity to samples at other locations. In the former case, it is important to obtain adequate coverage of covariate space (to assure parameter estimates required for prediction); in the latter case, we require adequate coverage of geographic space. Models can learn from both, and sampling design (e.g., Xia et al. 2006) and suppression schemes (Silberstein et al. 2007, Howard and Flikkema 2008) can exploit these relationships as basis for design of networks and algorithms.

Predictive distributions of data not yet collected, as opposed to raw data, can be more directly used in subsequent analyses, because they can be evenly distributed in space and time, and observation errors are already accommodated and enter as part of the prediction variances. By contrast, the corresponding raw data could require a sophisticated and time-consuming treatment by each new user having varying capacities to handle them appropriately, including proper interpretation of metadata on sensor and network problems. There are many advantages to a standard and transparent protocol for each type of error about which the developers have most insight.

The concern that predictive distributions are conditioned on assumptions is an important one. Yet this is true for any statistic, including a simple mean and

variance of observations, which would ignore instrument bias and errors that fluctuate in time. Reverting to raw data does not obviate the problem of data uncertainty, it just puts it off to the next stage of analysis, where it could be more difficult to accommodate. Ignoring uncertainty could mean that raw data misrepresent the processes of interest, to the extent that such biases and missing values exist.

Sensing networks that react to change in the environment (Batalin et al. 2005, Cardell-Oliver et al. 2005, Collins et al. 2006) represent an important first step in the direction we consider here. Ecologists have a long tradition of data modeling, once data are in hand. We simply consider how extending that approach to the data collection stage can help to maximize explanatory power for a given (minimum) expense, while improving access and interpretation for users with a range of objectives.

Acknowledgments

Funding for this study was supported by National Science Foundation grants IDEA-0308498, SEII 0430693, and DDDAS 0540347.

Literature Cited

Batalin, M., W. Kaiser, R. Pon, G. Sukhatme, G. Pottie, Y. Yu, J. Gordon, M. Rahimi, and D. Estrin. 2005. Task allocation for event-aware spatiotemporal sampling of environmental variables. IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005. IEEE, Washington, D.C., USA.

Cardell-Oliver, R., M. Kranz, K. Smettem, and K. Mayer. 2005. A reactive soil moisture sensor network: design and field evaluation. International Journal of Distributed Sensor Networks 1:149–162.

Cava, D., U. Giostra, M. Siqueira, and G. G. Katul. 2004. Organized motion and radiative perturbations in the nocturnal canopy sublayer above an even-aged pine forest. Boundary Layer Meteorology 112:129–157.

Collins, S. L., L. M. A. Bettencourt, A. Hagberg, R. F. Brown, D. I. Moore, G. Bonito, K. A. Delin, S. P. Jackson, D. W. Johnson, S. C. Burleigh, R. R. Woodrow, and J. M. McAuley. 2006. New opportunities in ecological sensing using wireless sensor networks. Frontiers in Ecology and the Environment 4:402–407.

Flikkema, P. G., P. K. Agarwal, J. S. Clark, C. Ellis, A. Gelfand, K. Munagala, and J. Yang. 2006. Model-driven dynamic control of embedded wireless sensor networks. Proceedings of the 6th International Conference on Computational Science, Workshop on Dynamic Data Driven Application Systems, Reading, UK.

Ganesan, D., A. Cerpa, W. Ye, Y. Yu, J. Zhao, and D. Estrin. 2004. Networking issues in wireless sensor networks. Journal of Parallel and Distributed Computing 64:799–814.

Howard, S. L., and P. Flikkema. 2008. Progressive joint coding, estimation and transmission censoring in energy-centric wireless data gathering networks. Proceedings 5th IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS 2008), Atlanta, Georgia, USA. IEEE, Washington, D.C., USA.

Lane, A. M. J. 1997. The U.K. environmental change network database: an integrated information resource for long-term monitoring and research. Journal of Environmental Management 51:87–105.

Martinez, K., J. K. Hart, and R. Ong. 2004. Environmental sensor networks. Computer 37:50–56.

McIntire, D., K. Ho, B. Yip, A. Singh, W. Wu, and W. J. Kaiser. 2006. The low power energy aware processing (LEAP) embedded networked sensor system. Proceedings Fifth International Conference on Information Processing in Sensor Networks (IPSN) 2006. Association for Computing Machinery, New York, New York, USA.

Naumburg, E., D. Ellsworth, and G. G. Katul. 2001. Modeling daily understory photosynthesis of species with differing photosynthetic light dynamics in ambient and elevated CO2. Oecologia 126:487–499.

Ni, K., N. Ramanathan, M. N. Chehade, L. Balzano, S. Nair, S. Zahedi, E. Kohler, G. Pottie, M. Hansen, and M. Srivastava. 2009. Sensor network data fault types. ACM Transactions Sensor Networks 5:1–29.

Oren, R., and D. Pataki. 2001. Transpiration in response to variation in microclimate and soil moisture in southeastern deciduous forests. Oecologia 127:549–559.

Porter, J., et al. 2005. Wireless sensor networks for ecology. BioScience 55:561–572.

Pottie, G. J. 2001. Wireless integrated network sensors (WINS): the web gets physical. National Academy of Engineering: The Bridge 31:22–27.

Pottie, G. J., and W. J. Kaiser. 2005. Principles of embedded networked systems design. Cambridge University Press, Cambridge, UK.

Richardson, A. D., D. Y. Hollinger, G. G. Burga, K. J. Davis, L. B. Flanagan, G. G. Katul, J. W. Munger, D. M. Ricciuto, P. C. Stoy, A. E. Suyker, S. B. Verma, and S. C. Wofsy. 2006. A multi-site analysis of uncertainty in tower-based measurements of carbon and energy fluxes. Agricultural and Forest Meteorology 136:1–18.

Silberstein, A., G. Puggioni, A. Gelfand, K. Munagala, and J. Yang. 2007. Suppression and failures in sensor networks: a Bayesian approach. Proceedings of the 33rd International Conference on Very Large Data Bases. Vienna, Austria.

Stoy, P., G. G. Katul, M. B. S. Siqueira, J. Y. Juang, K. A. Novick, J. M. Uebelherr, and R. Oren. 2006. An evaluation of models for partitioning eddy covariance-measured net ecosystem exchange into photosynthesis and respiration. Agricultural and Forest Meteorology 141:2–18.

Szewczyk, R., E. Osterweil, J. Polastre, M. Hamilton, A. Mainwaring, and D. Estrin. 2004. Habitat monitoring with sensor networks. Communications of the ACM 47:34–40.

Ward, E. J., R. Oren, B. D. Sigurdsson, P. G. Jarvis, and S. Linder. 2008. Fertilization effects on mean stomatal conductance are mediated through changes in the hydraulic attributes of mature Norway spruce trees. Tree Physiology 28:579–596.

White, M. A., and R. R. Nemani. 2006. Real-time monitoring and short-term forecasting of land surface phenology. Remote Sensing of Environment 104:43–49.

Xia, G., M. L. Miranda, and A. Gelfand. 2006. Approximately optimal spatial design approaches for environmental health data. Environmetrics 17:363–385.

Xiao, J.-J., A. Ribeiro, Z.-Q. Luo, and G. B. Giannakis. 2006. Distributed compression-estimation using wireless sensor networks. IEEE Signal Processing Magazine 23(4)::27–41.

Zhang, X., M. A. Freidl, and C. B. Schaff. 2006. Global vegetation phenology from moderate resolution imaging spectroradiometer (MODIS): evaluation of global patterns and comparison with in situ measurements. Journal of Geophysical Research 111:G04017.