From Data Reverence to Data Relevance: Model-Mediated Wireless Sensing of the Physical Environment

Paul G. Flikkema[1], Pankaj K. Agarwal[2], James S. Clark[2], Carla Ellis[2], Alan Gelfand[2], Kamesh Munagala[2], and Jun Yang[2]

[1] Northern Arizona University, Flagstaff AZ 86001 USA
[2] Duke University, Durham, NC USA

Abstract. Wireless sensor networks can be viewed as the integration of three subsystems: a low-impact *in situ* data acquisition and collection system, a system for inference of process models from observed data and *a priori* information, and a system that controls the observation and collection. Each of these systems is connected by feedforward and feedback signals from the others; moreover, each subsystem is formed from behavioral components that are distributed among the sensors and out-of-network computational resources. Crucially, the overall performance of the system is constrained by the costs of energy, time, and computational complexity. We are addressing these design issues in the context of monitoring forest environments with the objective of inferring ecosystem process models. We describe here our framework of treating data and models jointly, and its application to soil moisture processes.

Keywords: Data Reverence,Data Relevance, Wireless Sensing.

1 Introduction

All empirical science is based on measurements. We become familiar with these quantitative observations from an early age, and one indication of our comfort level with them is the catchphrase "ground truth". Yet one characteristic of the leading edge of discovery is the poor or unknown quality of measurements, since the instrumentation technology and the science progress simultaneously, taking turns pulling each other forward in incremental steps.

Wireless sensor networking is a new instrument technology for monitoring of a vast range of environmental and ecological variables, and is a particularly appropriate example of the interleaving of experiment and theory. There are major ecological research questions that must be treated across diverse scales of space and time, including the understanding of biodiversity and the effects on it of human activity, the dynamics of invasive species (Tilman 2003), and identification of the web of feedbacks between ecosystems and global climate change. Wireless sensor networks have great potential to provide the data to help answer these questions, but they are a new type of instrumentation with

Y. Shi et al. (Eds.): ICCS 2007, Part I, LNCS 4487, pp. 988–994, 2007.

substantial constraints: the usual problems of transducer noise, nonlinearities, calibration, and sensitivity to temperature and aging are compounded by numerous potential sensor and network failure modes and intrinsically unreliable multihop data transmissions.

Moreover, the entire measurement and networking enterprise is severely constrained by limited power and energy. There is substantial redundancy in data collected within wireless networks (Fig. 1). Yet the capacity to collect dense data when it provides valuable information is one of the key motivations for the technology. Clearly, there is need to control the measurement process with model-based evaluation of potential observations.

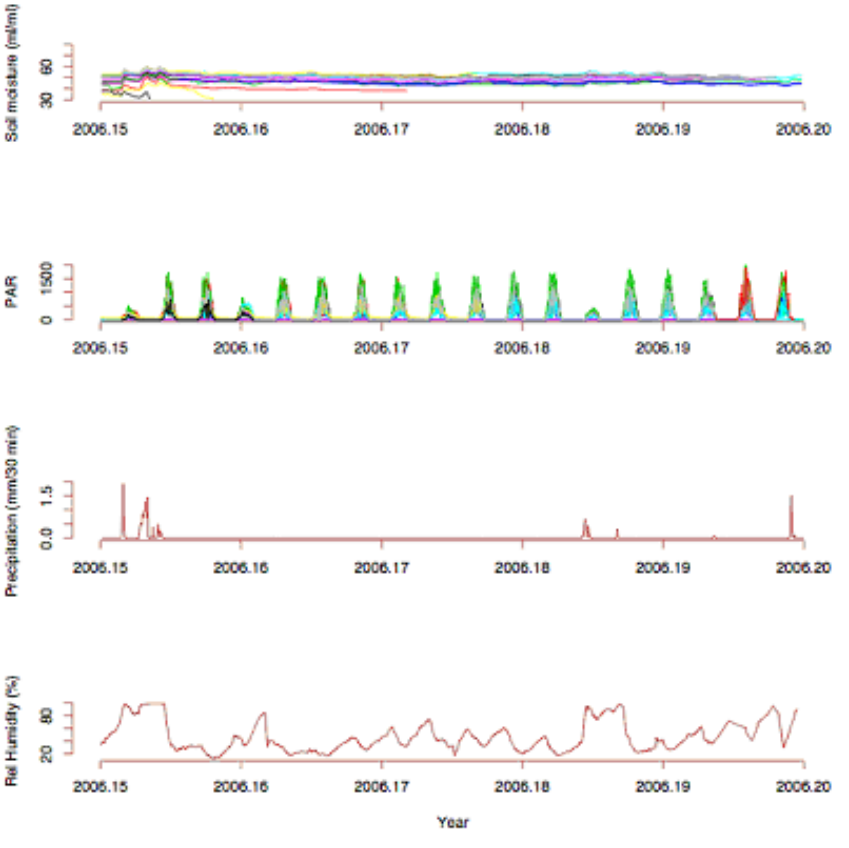

Fig. 1. Examples of four variables measured in the Duke Forest wireless sensor network showing different levels of redundancy at different scales

Measurements without underlying data and process models are of limited use in this endeavor. Indeed, most useful measurements, even when noiseless and unbiased, are still based on an underlying model, as in processing for sensors and satellite imagery (Clark et al. 2007). These often-implicit or even forgotten

models have limitations that are bound up in their associated data. For example, a fundamental operation in environmental monitoring is sampling in space and time, and one approach is to estimate temporal and spatial bandwidths to establish sampling rates. However, the usual frame of reference in this case is the classic Shannon sampling theorem, and the requirement of finite bandwidth in turn forces definition of the signal over all time, a clear model-based limitation.

The phenomena of interest in environmental monitoring are highly time-varying and non-stationary, and laced with measurement and model uncertainty. These factors are the key motivation for the application of the dynamic distributed data application systems (DDDAS) paradigm to wireless sensor networks (Flikkema et al. 2006). DDDA systems are characterized by the coupling of the concurrent processes of data collection and model inference, with feedbacks from one used to refine the other. The fact that resources—energetic and economic—are limited in wireless sensor networks is in some sense an opportunity. Rather than accept measurements as the gold standard, we should embrace the fact that both measurements and models can be rife with uncertainty, and then tackle the challenge of tracking and managing that uncertainty through all phases of the project: transducer and network design; data acquisition, transfer, and storage; model inference; and analysis and interpretation.

2 Dynamic Control of Network Activity

Looking at two extreme cases of data models—strong spatial correlation combined with weak local correlation and vice versa—can shed some light on the trade-offs in designing algorithms that steer network activity. First, consider the case when the monitored process is temporally white but spatially coherent. This could be due to an abrupt global (network-wide) change, such as the onset of a rainstorm in the monitoring of soil moisture. In this case, we need snapshots at the natural temporal sampling rate, but only from a few sensor nodes. Data of the needed fidelity can then be obtained using decentralized protocols, such as randomized protocols that are simple and robust (Flikkema 2006). Here, the fusion center or central server broadcasts a cue to the nodes in terms of activity probabilities. The polar opposite is when there is strong temporal coherence but the measurements are statistically independent in the spatial domain. One example of this is sunfleck processes in a forest stand with varying canopy density. Since most sensor nodes should report their measurements, but infrequently, localized temporal coding schemes can work well.

Our overall effort goes beyond data models to the steering of network activity driven by ecosystem process models, motivated by the fact that even though a measured process may have intrinsically strong dynamics (or high bandwidth), it may be driving an ecosystem process that is a low-pass filter, so that the original data stream is strongly redundant with respect to the model of interest. Our approach is to move toward higher-level modeling that reveals what data is important.

A common criticism might arise here: what if the model is wrong? First, given the imprecision and unreliability of data, there is no a priori reason to favor data. For example, we often reject outliers in data preprocessing, which relies on an implicit "reasonableness" model. Yet an outlier could be vital information. Thus any scheme must dynamically allocate confidence between the model and the incoming data. By using a Bayesian approach, this allocation can be made in a principled, quantitative manner (Clark 2005, Clark 2007, MacKay 2003).

Any algorithm that uses model-steered sampling and reporting (rather than resorting to fixed-rate sampling at the some maximum rate) will make errors with a non-zero probability. To mitigate these errors, our strategy is based concurrent execution of the same models in the fusion center as in individual sensors. Using this knowledge, the fusion center can estimate unreported measurements; the reliability of these estimates is determined by the allowed departure of the predicted value from the true value known by the sensing node. The fusion center can also run more sophisticated simulation-based models that would be infeasible in the resource-constrained sensors, and use the results to broadcast model-parameter updates.

Clearly, a missing report could be due to a communication or processing error rather than a decision by the sensor. By carefully controlling redundancy within the Bayesian inference framework, which incorporates models for both dynamic reporting and failure statistics (Silberstein et al. 2007), it become possible to infer not only data and process models, but node- and network-level failure modes as well. Finally, in our experiments, each sensor node archives its locally acquired data in non-volatile memory, allowing collection of reference data sets for analysis.

3 Example: Soil Moisture Processes

Soil moisture is a critical ecosystem variable since it places a limit on the rate of photosynthesis and hence plant growth. It is a process parameterized by soil type and local topography and driven by precipitation, surface runoff, evapotranspiration, and subsurface drainage processes. Because it is highly non-linear, it is much more accessible to Bayesian approaches than ad hoc inverse-modeling techniques. Bayesian techniques permit integration of process noise that characterizes our level of confidence in the model. In practice, it may more productive to use a simple model with fewer state variables and process noise instead of a model of higher dimension with poorly known sensitivity to parameter variations.

Once the model is obtained (for example, using training data either from archival data or a "shake-out" interval in the field), the inferred parameters can then be distributed to the sensor nodes. The nodes then use the model as a predictor for new measurements based on the past. The observation is transmitted only when the discrepancy between the measurement and the predicted value exceeds a threshold (again known to both the sensor and the fusion center). Finally, the model(s) at the fusion center are used to recover the unreported changes.

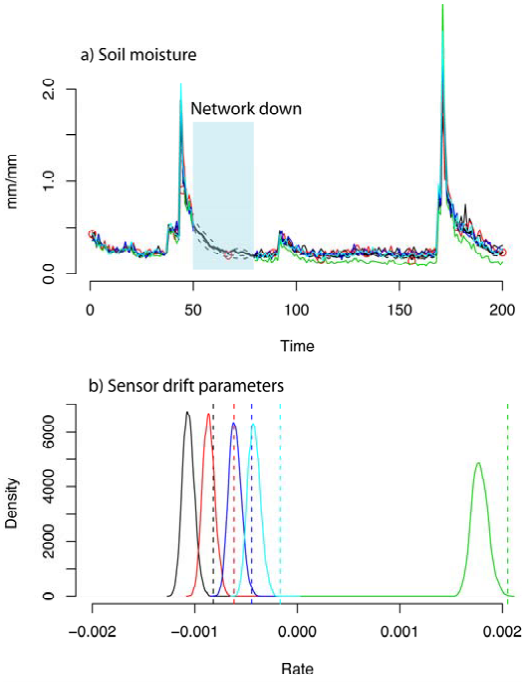

Fig. 2. a) Simulated soil moisture data (solid line) and simulated observations (colored lines) from sensors that drift. b) Posterior estimates of parameter drift become part of the model that is used to determine which observations to collect (Fig. 3).

In the simulation results shown in Figure 2 (Clark et al. 2007), the underlying soil moisture is shown as a solid black line. Here we use a purely temporal model in each sensor node. 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 assume that all sensors are down for a period of time. The 95% predictive intervals for soil moisture (dashed lines) 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 (Figure 2b), but these parameters are of limited interest, and have limited impact on predictive capacity (Figure 2a). The impact on reporting rate and associated energy usage is substantial as well (Clark et al. 2007).

Our strategy is to incorporate dynamic reporting starting with simple, local models in an existing wireless sensor network architecture (Yang 2005). As shown by the soil moisture example, even purely temporal models can have a significant impact. From a research standpoint, it will be useful to first determine the effectiveness of dynamic reporting driven by a local change-based model where a node reports an observation only if it has changed from the previously reported observation by a specified absolute amount. This is simple to implement and requires a negligible increase in processing time and energy. In general, local

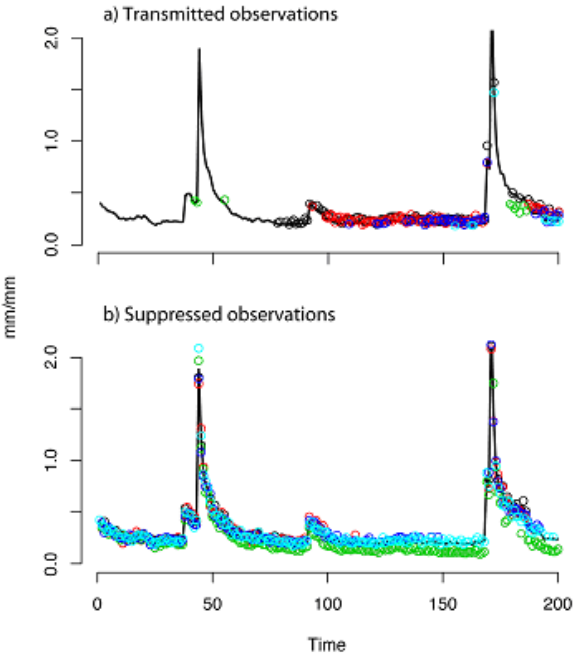

Fig. 3. A simple process model, with field capacity and wilting point, together with a data model that accommodated parameter drift (Fig. 2b) allows for transmission of only a fraction of the data (solid dots in (a)). Far more of the measurements are suppressed (b), because they can be predicted.

models have the advantage of not relying on collaboration with other sensor nodes and its associated energy cost of communication.

What about the problem of applying one model for data collection and another for modeling those data in the future? It is important to select models for data collection that emphasize data predictability, rather than specific parameters to be estimated. For example, wilting point and field capacity are factors that make soil moisture highly predictable, their effects being evident in Figures 1 and 2. By combining a process model that includes just a few parameters that describe the effect of field capacity and wilting point and a data model that includes sensor error, the full time series can be reconstructed based on a relatively small number of observations (Figure 3)(Clark et al. 2007).

4 Looking Ahead

Researchers tend to make an observation, find the most likely value, and then treat it as deterministic in all subsequent work, with uncertainty captured only in process modeling. We have tried to make the case here for a more holistic approach that captures uncertainty in both data and models, and uses a framework to monitor and manage that uncertainty. As wireless sensor network

deployments become larger and more numerous, researchers in ecology and the environmental sciences will become inundated with massive, unwieldy datasets filled with numerous flaws and artifacts. Our belief is that much of this data may be redundant, and that many of the blemishes may be irrelevant from the perspective of inferring predictive models of complex, multidimensional ecosystems processes. Since the datasets will consume a great deal of time and effort to document, characterize, and manage, we think that that the time for model-mediated sensing has arrived.

References

1. NEON: Addressing the Nation's Environmental Challenges. Committee on the National Ecological Observatory Network (G. David Tilman, Chair), National Research Council. 2002. ISBN: 0-309-09078-4.
2. Clark, J.S. Why environmental scientists are becoming Bayesians. Ecol. Lett. 8:2-14, 2005.
3. Clark, J.S. *Models for Ecological Data: An Introduction.* Princeton University Press, 2007.
4. Clark, J.S., Agarwal, P., Bell, D., Ellis, C., Flikkema, P., Gelfand, A., Katul, G., Munagala, K., Puggioni, G., Silberstein, A., and Yang, J. Getting what we need from wireless sensor networks: a role for inferential ecosystem models. 2007 (in preparation).
5. Flikkema, P. The precision and energetic cost of snapshot estimates in wireless sensor networks. Proc. IEEE Symposium on Computing and Communications (ISCC 2006), Pula-Cagliari, Italy, June 2006.
6. Flikkema, P., Agarwal, P., Clark, J.S., Ellis, C., Gelfand, A., Munagala, K., and Yang, J. Model-driven dynamic control of embedded wireless sensor networks. Workshop on Dynamic Data Driven Application Systems, International Conference on Computational Science (ICCS 2006), Reading, UK, May 2006.
7. MacKay, D.J.C. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003.
8. Silberstein, A., Braynard, R., Filpus, G., Puggioni, G., Gelfand, A., Munagala, K., and Yang, J. Data-driven processing in sensor networks. Proc. 3rd Biennial Conference on Innovative Data Systems Research (CIDR '07), Asilomar, California, USA, January 2007.
9. Yang, Z., et al. WiSARDNet: A system solution for high performance in situ environmental monitoring. Second International Workshop on Networked Sensor Systems (INSS 2005), San Diego, 2005.