# Intelligent Systems

## Terrestrial Observation and Prediction Using Remote Sensing Data

Joseph C. Coughlan
Mail Stop 269-3, Intelligent Systems Division
NASA Ames Research Center
Moffett Field, California, USA.
joseph.c.coughlan@nasa.gov

*Abstract*—**NASA has made science and technology investments to better utilize its large space-borne remote sensing data holdings of the Earth. With the launch of Terra, NASA created a data-rich environment where the challenge is to fully utilize the data collected from EOS however, despite unprecedented amounts of observed data, there is a need for increasing the frequency, resolution, and diversity of observations. Current terrestrial models that use remote sensing data were constructed in a relatively data and compute limited era and do not take full advantage of on-line learning methods and assimilation techniques that can exploit these data. NASA has invested in visualization, data mining and knowledge discovery methods which have facilitated data exploitation, but these methods are insufficient for improving Earth science models that have extensive background knowledge nor do these methods refine understanding of complex processes. Investing in interdisciplinary teams that include computational scientists can lead to new models and systems for online operation and analysis of data that can autonomously improve in prediction skill over time.**

***Keywords-component; intelligent systems; remote sensing; prediction; model; NASA***

## I. Introduction

NASA research focuses on providing data and information derived from space-based remote sensing systems to answer social-economically important questions about the Earth. Living systems are dynamic and require frequent repeat observations at both moderate (10-1) km and high spatial resolutions (1km -30m) as well as complementary airborne and *in situ* observations. Intensive field campaigns, process studies, fundamental research, data and information systems, and modeling are all essential for interpreting satellite observations and providing answers to science questions. Effective use of all these data requires extensive reformulation of current techniques and models

The ultimate goal is to project future conditions and trends for ecosystems and the global carbon cycle as well as providing key inputs for climate models, like future atmospheric $CO_2$ and $CH_4$ concentrations and representations of key ecosystem and carbon cycle process controls on the climate system. Three currently independent project elements must be integrated to achieve this goal: i) Modeling, ii) intelligent systems, and iii) online analysis.

## II. Background

### A. Modeling

Scientific models are constructed by hand from a mix of first-principle based understanding and computational approximations, possibly non-parametric, built from analysis of synoptic and time-series data. Modeling requires defining a system; its boundary, the internal sub-components, and the interactions between the components and with the external environment. The boundary delineates the system and defines the interfaces between the system and the environment. The model requires initialization data, input data that define the external environment influences at each time step and internal state calculations quantifying the system components. Confidence in the model is determined by analysis of how its process sub-models are coupled, how they mimic subsystem behavior, and how well simulated dynamics compare to independent observations, whenever such comparisons are practical.

Model construction has historically been a divide and conquer process, defining components only when necessary to i) reduce error, ii) add representations to answer questions asked about the system or iii) to create intermediate calculations for validation with observations. Error is assumed to be due to missing or malformed system representations that when addressed, improve the model's fidelity. Questions about a how system functions can require analysis of specific calculations and state variables. For example, understanding the impact of forest canopy density on forest hydrology across the USA requires modeling snow pack dynamics since it is critical to the hydrologic cycle and while observable initializing a plot-sized simulation, snow pack difficult to observed for regional applications and must be modeled. Adding components that generate intermediate calculations can help validate a model and allow for comparisons to observable conditions. A useful intermediate measure for forest model validation is pre-dawn leaf water potential, an indictor of soil moisture, canopy water content, leaf stomatal conductance, and plant physiological health.

As our technical sophistication evolves so too can models improve by developing better computational representations

NASA Intelligent Systems Project

and model implementations. Terrestrial modeling has not advanced as far as computing technology has advanced over the past 20 years. Intel's CPUs clock speeds have increased four orders of magnitude; from a 3.5 MHz 8008 to 3.5 GHz 8088 compatible processor, yet the basic design of terrestrial models have been is relatively conservative over this period. Model improvements can be measured by degree of fidelity to first principles, improved mimicry of observed behavior, the degree of previous modeling knowledge used and how easily current hypotheses of system behavior can be incorporated in models. Most notable is how few of the observational data collected have been incorporated into models compared to what is possible given the magnitude of technological advancement due to Moore's Law.

### B. *Intellgent Systems*

The computational field most familiar to Earth scientists is high end computing but there are a collection of important disciplines collectively referred to as "intelligent systems" that have great promise to help modeling. These are i) data mining, ii) knowledge discovery in databases and iii) machine learning. They overlap but can be distinguished by the progression general questions each answers: i) "What sub-structure is in the data?" ii) "why and how 'it' is happening?" iii) "Where and when will 'it' occur again?"

When used as out-of-the-box tools, these methods often answer "Sesame Street" questions, "Which of these things is not like the others?" While discriminating non-trivial sub-structure in large datasets (i.e. anomalies,.patterns etc.) is very useful, that type information is not directly communicable to model refinement or to enhancing scientific understanding. For example, a system can be made to scan remotely sensed data to recognize volcanoes on Venus or find and label patterns of fire risk in images of terrestrial systems. The automated identification of these structures reduces data complexity but is not a communicable result for refining a model. These methods also have application when searching model output, which can be enormous, during parameter searches and sensitivity analysis. Here the question is "what sub-structure is in the data?" and answering it can help identify outliners or system thresholds. Answers to these questions do not produce results that improve model design without extensive human interpretation, experimentation, and intervention.

Machine learning (ML) has been used to improve models by mimicking the behavior of more complex first principle based models. A non-parametric machine learning algorithm has been trained with a complex parametric model and can then efficiently reproduce results with one or more orders of magnitude less overhead. These ML solutions are suitable for operational use but do not increase understanding. Examples in use include replacing look-up tables with faster ML approximations in EOS product processing and replacing the long wave radiation sub-model in an NCAR climate model. Here the question answered is "Where and when will 'it' occur again?" The ML method can correctly predict the output when given an input vector.

There exist machine driven methods that can build parametric models from data but these techniques are menu driven, do not account for uncertainty, and can only reproduce simple, deterministic parametric models with fitted coefficients.

### C. *On-line Analysis*

Currently the most advanced terrestrial biophysical variable estimated from an online satellite data stream is the MODIS MOD-17 product, an 8-day composite of daily NPP estimates in global coverage. This 1km product is often spatially aggregated to reduce data volume for global monitoring. The basis for MOD-17 is a photosynthetic efficiency model where the efficiency factor, $\varepsilon$, is dynamically calibrated by a high fidelity ecosystem model that contains detailed soil and plant physiology sub-models.

EOS products are produced as quickly as possible and there is great importance placed on doing the calculations rapidly however, there is no real-time constraint for product generation and it is possible for the product processing to lag current time by weeks.

The EOS product is updated in versioned releases and there is no capability to improve or update model parameters using comparisons to observed data. This design is consistent with the entire EOS algorithm development philosophy.

The validation of the MOD-17 algorithm relies heavily on independent estimates made at large (km) "footprint" estimates of surface fluxes measured by eddy towers and during unique, field campaigns that benchmark key ecosystem types (FIFE, OTTR, BOREAS, & LBA). As knowledge is gain, models are inspected, updated, and engineered for possible updating in the EOS production system. This is a labor-intensive process and requires a project office and handcrafted modifications.

## III. Recommendations

The challenge is to combine i) modeling, ii) intelligent systems and iii) online analysis to automate the scientific process and incorporate data into so they can be improved and refined as data are collected. This challenge requires data analysis and summarization methods that can be incorporated with existing knowledge into models that autonomously improve as data are collected and analyzed. A critical step to achieve this goal is interdisciplinary research between the computational and natural sciences that support high risk, development of new model and methods that can fully assimilate and ingest data from a variety of sources.

### A. *Modeling*

There is a large enough data record and the prospects of future data collection to warrant research to construct data driven terrestrial models that are parametric, use first principles where possible, and are designed for on-line refinement. The model structure can be modified by hand if hypothesized sub-components' are not necessary or if new questions are being asked. Previous knowledge learned about parameter values can be reused in the new models.

This is shift away from the historical, handcrafted models built in data poor environments to new models designed to assimilate and improve as large volumes of data are collected and ingested. If validation data, which independently estimate the surface biophysical parameters, is made available in the modeling architecture, then the model can learn from comparisons between observed and predicted and update the initial model parameters to improve on prediction while adhering to the physical representations in the parametric model's structure.

### *B. Intellgent Systems*

Data mining and KDD investments should not be treated as after-the-fact analysis tools but developed as part of these new models and designed to work within the model framework. Treating these methods as general-purpose tools and capabilities available in commercial software will not improve the modeling process. These methods can be designed to summarize and mine data to answer specific questions and catalogue features.

Mining can automatically generate summaries data by extracting higher order features within in the data such as tracking surface clusters of "fire risk". These clusters can trigger high resolution simulations, observations and forecasts for those at risk regions. The fire-risk clusters are an example of a higher-level object that summarize data and they can be efficiently tracked and analyzed as dimensionless objects with a small set of descriptive attributes. These objects have meaning and are less complex than the raw data. A non-parametric prediction model can be built from a catalogue of these high-level features. *i.e.* "How often is a fire observed in fire risk clusters of size X over time duration Y?"

KDD methods can be deployed in decision support systems to apply these models for case specific applications. KDD systems are currently visualization based systems but can be extended and closely coupled to models and analysis methods in order to help users interactively refine data clusters, label data for sharing and/or produce first cut computational approximations describing the causes (precursor variables and values) of these features.

### *C. On-line Analysis*

An on-line monitoring and prediction capability should focus on making useful observations and near-term predictions available as soon as possible to meet time-critical user needs. An on-line capability should not compromise timeliness for completeness. If accuracy is needed for trending analysis then a second calculation can produce are more complete result for that purpose. This second calculation can be done offline and not compromise real-time, online performance.

Data streams collected for product validation should be made available for ingest into satellite-based product generation and contingencies established for coping with instances of missing or corrupt data. The system can automatically document the data heritage and methods used to produce the data product.

As observation data are ingested, model results can be compared to on-line validation data and discrepancies used to update model parameters and improve future estimates. Over time, model parameters should converge and product errors shrink in comparison to ground observations. Automating this process will increase model performance, greatly reduce the cost and complexity of using data in models and increase the effective use of data collected by NASA's space borne assets.

## IV. Summary

NASA's ultimate goal is to project future conditions and trends for ecosystems and the global carbon cycle as well as providing key inputs for climate models. Three currently independent facets of research must be integrated to improve our predictive capability for terrestrial ecosystems: i) Modeling ii) intelligent systems and iii) online analysis. These elements are interdependent and when developed independently do not integrate well into an monitoring and prediction system. When integrated and designed to accommodate large data volumes, these technologies can full utilize the large volume of complex data by leveraging future advances in computational systems and algorithm research. When left independently, there will be a mismatch between the technology capabilities and science requirements.

## Acknowledgment

I gratefully acknowledge Dr. Dogan Timucin and Dr. Rama Nemani for helping in formulating ideas, discussions, and comments on early drafts of this conference paper..