2.1

A DATA MINING APPROACH TO SOIL TEMPERATURE AND MOISTURE PREDICTION

William Myers*, Seth Linden, and Gerry Wiener
National Center for Atmospheric Research, Boulder, CO

1. INTRODUCTION

Both weather and soil conditions are important factors in the agricultura* l decision making process. For example, phenological pest models that predict the evolution of an organism's life and its impact on crops are based on the temperature of its environment. Accurate soil state predictions can dramatically improve the decision making process of the agricultural community.

Typically, soil temperature and moisture forecasting has been performed using a land-surface model (LSM). This is a physically-based approach that models heat transfer and moisture flow between the atmosphere and the soil subsurface. It is usually initialized with current subsurface and atmospheric conditions. This NASA-funded project uses such a physical modeling system, the High Resolution Land Data Assimilation System (HRLDAS) and the Noah LSM, to model the evolution of the soil state. The Noah LSM is a widely used land surface model which is included in the WRF as part of NCEP's operational NAM. It requires a large number of parameters to effectively simulate the energy transfer for different land use, soil types, and various vegetation states.

This paper describes another method to predict soil temperature and moisture prediction. In this case, a machine learning approach using a regression tree algorithm (Cubist) was developed to predict the future soil state based upon current soil state and atmospheric conditions. Soil state observations and the following hour's observed weather were used as predictors in the regression tree and the next hour's soil state was set to be the target. In a regression tree, the data is analyzed and rules are developed that break the data down into a number of different cases, producing a decision tree. A regression equation is then generated at each leaf node. In essence, a linear approximation to the highly non-linear LSM processes is produced for each case.

An example rule and associated regression would be:

if dsw <= 0.09 and ST5_prev > 12.05
ST5_curr = -0.211 + 0.3165 dsw +
0.83 ST5_prev + 0.13 ST10_prev +
0.02 AirT + 0.02 TD

That is, if it is night (no downward shortwave solar radiation) and the last hour's 5 cm temperature was above 12ºC, apply the regression equation (shown above) which combines downward solar radiation, the previous hour's 5 cm soil temperature, the previous hour's 10 cm soil temperature, the mean air temperature in the current hour, and the mean dew point temperature in the current hour.

2. METHOD

The machine learning technique is dependent upon soil observations. There are a limited number of soil observation sites in the U.S. The Soil Climate Analysis Network (SCAN), part of the USDA's Natural Resources Conservation Service, was chosen for use in this study due to its broad geographic extent (within the continental U.S.) and its uniform set of observed atmospheric and soil state variables. The network is reasonably well maintained and its data are of fairly good quality. These sites usually measure soil temperature and moisture at depths which correspond directly to the HRLDAS/Noah LSM's node depths. This simplifies the verification process. SCAN sites with fairly complete observational histories were selected from within this project's agriculturally-oriented domain (east of the Rockies).

Land-surface modelers were consulted to determine the critical atmospheric forcing variables that drive the soil state evolution. Cubist was provided with 2 years (2005 and 2006) of training data for each SCAN site. Using this data, regression trees were developed specific to each SCAN site. Separate regression trees were developed to predict soil moisture and soil temperature at 5 cm, 20 cm, 50 cm, and 1m below ground. Effectively, eight separate regression trees were produced for each site. By providing the current soil temperature and moisture state, along with the predicted atmospheric conditions, these eight sets of rules and regressions could be applied to predict the soil state one hour into the future. A 48 hour soil forecast could then be generated by iteratively

* *Corresponding author address:* William Myers, NCAR, 3450 Mitchell Lane, Boulder, CO 80302; e-mail: myers@ucar.edu

using the predicted soil state with each hour's weather forecast.

These regression trees were applied to the 2007 growing season (April-June). For every day within this period, a 48 hour soil forecast was produced at each site using the same weather data used to drive the physically-based model. The machine-learning forecasts were compared to forecasts generated by the HRLDAS/Noah system using mean absolute error (MAE) and bias calculated at each site over the length of the growing season.

3. RESULTS

For soil temperature, the regression tree forecasts were better at nearly all the sites and depths. At 5 cm (the most important for most agricultural applications) and 20 cm, the forecasts were clearly better at 24 of the 29 forecast points. At 50 cm, the data mining forecasts were better at 28 of the 29 points. There were only two sites where the physical model outperformed the data mining approach at more than one depth.

Order statistics for the MAEs at each site showed that the data mining results were significantly better across the board. The extremes and the quartiles of the 5 cm temperature forecast MAEs for the physical model were over 30% higher than those of the regression tree approach. At 20 cm, the soil-temperature forecast errors were more than 70% worse for the physical model. At 50 cm, the errors were more than 2.5 times larger for the physical model. The difference is largely due to a significant cold bias in the physical model. This is a known problem in the HRLDAS/Noah LSM model and is currently being addressed by the NCAR land surface modeling team. The soil temperature forecast MAEs for the data mining and physical models are shown in Tables 1-3.

The soil moisture forecasts from the physical model are a challenge during the growing season. The vegetation state is critical to correctly model the transfer of water between subsurface nodes. The vegetation type and state used by the physical model in this experiment was based on spatially and temporally coarse climatological data. It is not surprising then, that during this season of rapidly changing vegetation, the physical model's soil moisture errors are significantly higher than those of the data mining approach. These errors are summarized in Tables 4-6 below.

	Minimum	1st Quartile	Median	3rd Quartile	Maximum
Data Mining	0.53	0.68	1.02	1.28	2.67
HRLDAS	0.85	1.12	1.36	1.79	3.52

Table 1: Summary of the mean absolute errors for the 5 cm soil temperature forecast for hours 0-60 for the 29 SCAN sites in degC over the 2007 growing season (April-June)..

	Minimum	1st Quartile	Median	3rd Quartile	Maximum
Data Mining	0.29	0.36	0.60	0.82	2.05
HRLDAS	0.58	0.86	1.17	1.55	3.46

Table 2: Summary of the mean absolute errors for the 20 cm soil temperature forecast for hours 0-60 for the 29 SCAN sites in degC, Apr-June 2007.

	Minimum	1st Quartile	Median	3rd Quartile	Maximum
Data Mining	0.13	0.24	0.30	0.40	1.19
HRLDAS	0.33	1.00	1.32	1.79	4.06

Table 3: Summary of the mean absolute errors for the 50 cm soil temperature forecast for hours 0-60 for the 29 SCAN sites in degC, Apr-June 2007.

	Minimum	1st Quartile	Median	3rd Quartile	Maximum
Data Mining	0.48	1.30	1.71	2.10	4.02
HRLDAS	2.76	5.95	8.08	10.50	17.22

Table 4: Summary of the mean absolute errors for the 5 cm soil moisture forecast for hours 0-60 for the 29 SCAN sites in percentile, Apr-June 2007.

	Minimum	1st Quartile	Median	3rd Quartile	Maximum
Data Mining	0.15	0.71	1.30	1.73	4.88
HRLDAS	3.89	6.77	13.54	17.99	26.02

Table 5: Summary of the mean absolute errors for the 20 cm soil moisture forecast for hours 0-60 for the 29 SCAN sites in percentile, Apr-June 2007.

	Minimum	1st Quartile	Median	3rd Quartile	Maximum
Data Mining	0.08	0.28	0.75	1.40	10.47
HRLDAS	1.73	4.56	8.29	11.76	22.77

Table 6: Summary of the mean absolute errors for the 50 cm soil moisture forecast for hours 0-60 for the 29 SCAN sites in percentile, Apr-June 2007.

The physical model has a significant dry bias at all depths. This is a difficult issue to correct. HRLDAS has more subsurface nodes than observational depths. Soil temperature and moisture observations can be used to initialize the physical model at some depths, but the others cannot be directly set. This can lead to a significant shock due to slightly out of balance initial conditions, a bias early in the forecast, and the development of large forecast errors in future lead times. If observation data is not used to initialize the soil state, then previously generated model results can be used as pseudo-observations for initialization. However, over time the system may drift from reality. Even with use of observations it is difficult to eliminate the model's bias since the solution always trends back to a steady state based on the model's physics. The results for soil moisture prediction are shown in tables 4-6 and the results indicate much better performance for the data mining approach.

4. CONCLUSIONS

At soil observation sites, the data mining approach seems to work much better than the physical model. Assumptions in the parameterization must be made in order for the physical model to work over a large domain. The physical model's parameters could be tuned for that site; however, this would be a challenging multidimensional minimization problem for a highly non-linear model. In practice, this does not seem to be nearly as straightforward as the data mining approach. One attractive aspect of this machine learning approach is that it requires little parameter adjustment by the user.

The data mining approach is dependent upon a full set of observations. At locations where there is little or no observational history, this approach cannot easily be used. It would be worthwhile to develop regression trees for each land use and soil type pair and then apply those rules/regressions at similar non-observational locations. However, in the USGS data set, there are 27 land use types and 19 soil types. Unfortunately, there are so few soil observation sites compared to the number of land use and soil type combinations that this is not practical.

5. FUTURE WORK

Vegetation attributes are critical to soil temperature and moisture prediction. HRLDAS uses climatological vegetation data. The data mining approach does not explicitly take the vegetation state into account; however Cubist may be able to infer something about the vegetation state in the rules development by considering the month of the observation. A major goal of this project is to evaluate whether NASA MODIS satellite data can be used to provide better estimates of the current vegetation state and thus improve the soil temperature and moisture forecasts from HRLDAS. The development team is currently working to incorporate the MODIS data into both the data mining approach and the physical model. An evaluation of the impact of the MODIS data should be available by early next year.

It is also possible that regression tree rules could be generated by considering all the observational histories as one data set. By including all land use and soil types in the training data set, it may be possible to produce rules that work reasonably well everywhere. The rule generation process would hopefully distinguish when it was appropriate to make rules based on the land and soil characteristics. It is hypothesized that these forecasts would not be better than the regression trees tuned to a specific site; however, they may be competitive with the physical model and applicable at non-observing sites.

6. REFERENCES

6.1 Project Overview

Myers,W., F. Chen, 2008: Application of Atmospheric and Land Data Assimilation Systems to an Agricultural Decision Support System. 2007 AMS Conference on Agriculture and Forestry, Orlando, FL.

6.2 HRLDAS and Noah LSM

Chen,F., K. Manning, M. LeMone, S. Trier, J. Alfieri, R. Roberts, M. Tewari, D. Niyogi, T. Horst, S. Oncley, J. Basara, and P. Blanken, 2005: Description and Evaluation of the Characteristics of the NCAR High-Resolution Land Data Assimilation System, Vol.46, pp. 694-713

7. ACKNOWLEDGEMENTS

This research is sponsored by the NASA Earth Observing System, in response to NRA #NNH05ZDA001N, Research Opportunities in Space and Earth Sciences (ROSES-2005). The material within is based upon work supported by NASA under award No. NNA06CN03A titled “A Soil Temperature and Moisture Decision Support System for Agriculture”. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Aeronautics and Space Administration.

This project would like to acknowledge the data sets and help provided by the USDA Natural Resource Conservation Service (NRCS) who manage the SCAN data.

Cubist is a data mining software system available at www.rulequest.com.