THE MURI UNCERTAINTY MONITOR (MUM)

David W. Jones*
Applied Physics Laboratory, University of Washington
Seattle, WA

Susan Joslyn
Department of Psychology, University of Washington
Seattle, WA

1. INTRODUCTION

Modern military weather forecasters rely heavily on numerical models (Jones et al., 2002) to produce their forecasts. The reliability of this information is variable such that forecasters must also consider the amount of uncertainty inherent in the models' information. The steps involved in conducting a thorough evaluation of model uncertainty are time consuming. Forecasters must answer the following questions: How accurate have the models been over the past few days? How do the model initializations compare to the observational data? How uncertain are the current model predictions? Because military forecasting is often done under time pressure, the amount of uncertainty evaluation that can be conducted is often limited. In addition, the optimal means to convey this information to the end user is not well understood. Nevertheless, the level of uncertainty in a forecast can be a crucial factor in tactical decisions.

The design of the MUM system described below resulted from field observations and task analyses of operational Navy forecasters. MUM supports the process of uncertainty evaluation by taking on those steps that are computationally demanding for humans. This effort is part of a Department of Defense Multi-disciplinary University Research Initiative (MURI) on statistical and cognitive approaches to visualizing uncertainty in mesoscale meteorology, and is being conducted at the University of Washington in Seattle.

2. METHOD

Two studies were conducted to gain a better understanding of how uncertainty evaluation is conducted by operational Navy forecasters and how uncertainty fits into the forecasting process.

2.1 Verbal protocol analysis

We began with a cognitive task analysis (CTA) of operational forecasters at the Naval Pacific Meteorology and Oceanography Facility (NPMOF), Whidbey Island, WA. We observed four forecasters as they produced a Terminal Aerodrome Forecast (TAF), which is produced every six hours and amended when necessary. Not only is the TAF written under time pressure, but forecasters are also responsible for simultaneous duties. They respond to numerous requests in person and over the phone that necessarily interrupt the forecasting process.

For this study, the forecasters were instructed to verbalize their thoughts as they produced the TAF. We made audio recordings of their verbalization as well as video recordings to their computer screens. The auditory recordings were transcribed and broken down into individual numbered statements. Each statement was then coded and organized under goals.

For the most part the forecasters gathered information, primarily numerical models from the Web, early in the process. One forecaster delayed information gathering until he began to write his TAF, suggesting that he was avoiding maintaining a large memory load for the duration of the process. We noted that few forecasters had a detectable stage at which they built a complex mental model of the current atmospheric conditions (Trafton et al., 2000). Instead, most forecasters relied upon rule-of-thumb forecasting, applying standardized general rules to the current situation to derive the forecast (e.g., If a system is coming into the coast, strong southerly winds should be forecasted over Whidbey Island).

All forecasters made some effort to evaluate model uncertainty and talked about specific strategies for doing so. These were mostly complex mental comparisons of model output to other sources of information. Forecasters also discussed model biases, the quality of the initialization, and adjustments they would make to the model prediction to arrive at their own forecast.

Although some forecasters evaluated a number of different numerical models and compared them to various information sources, others focused on specific models and made only one or two comparisons. Forecasters tended to avoid head-to-head comparisons between models and complex quantitative evaluations. They rarely evaluated recent model performance and we did not observe them using either probability or ensemble products.

2.2 Questionnaire study

Our results were confirmed in a questionnaire study in which forecasters filled out a survey asking about model evaluation techniques after each TAF they produced. They also rated model performance. We expected to see the forecasting process altered when the models were judged to be performing poorly. Instead, forecasters appeared to have a fairly set forecasting routine, relying upon the same information sources and evaluation techniques from one TAF to the next.

3. FINDINGS FROM THE TASK ANALYSIS

From these two studies we learned that forecasters are concerned with model uncertainty and that they evaluated it on every TAF we observed. Naval forecasters clearly believe this is an important step, suggesting they will likely use products designed to facilitate this effort.

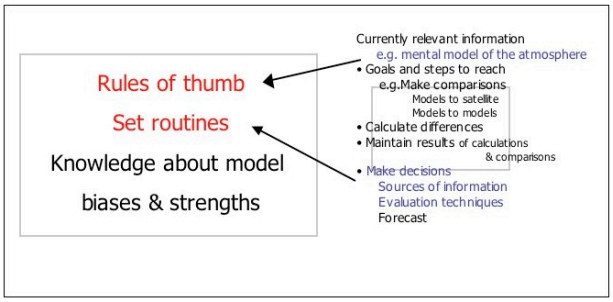

Figure 1. Example of forecaster cognitive offloading from short-term memory (right box) to long-term memory (left box).

Nonetheless, naval forecasters tended to avoid some procedures. We speculate that the avoided procedures are likely those placing the greatest demand on working memory. It has long been known that working memory, roughly synonymous with conscious level processing, is severely limited (Miller, 1956). Moreover, working memory capacity is functionally decreased by time pressure (Edland & Svenson, 1993), and interruptions (Rogers & Monsell, 1995), which are common in military forecasting. Tasks such as making complex mental comparisons, creating mental models of the current atmospheric conditions, and deciding which evaluation techniques are appropriate to a specific forecast draw heavily on this limited resource.

The forecasters we observed tended to avoid such tasks and relied instead upon approaches that tapped long-term memory (Figure 1). These are the solutions that can be memorized and applied with little adjustment to the current situation, such as rules of thumb or set routines. Although this approach alleviates working memory load, it may be at the cost of flexibility and thoroughness.

From the analysis of coded statements and in-depth review of the tools and information sources used by the forecasters we observed, we noted the following issues:

Forecasters are aware of model uncertainty and they attempt to estimate it by synoptic scale pattern matching and comparison of specific values.

Uncertainty evaluation is streamlined in response to time pressure and experience level.

Their primary information sources are numerical models. They receive most information via the web and rely on a small subset of available information. Despite this limited data set, they spend significant time navigating between info sources.

They have limited tools available to assess model uncertainty.

They are unsure how and when to use ensemble and other types of uncertainty products.

4. THE MUM SYSTEM ARCHITECTURE

A key objective of the UW MURI is to develop and test new visualizations of uncertainty information based on the UW Short Range Ensemble Forecast (SREF), Mass & Grimmit (2002). We felt that merely creating new products and then expecting overloaded forecasters to use them would yield disappointing results. What is needed is a framework that addresses all aspects of uncertainty that typically confront forecasters. They need a computer interface that presents only the essential information, yet allows for knowledge discovery when time or operational necessity requires it. We have developed the MUM as a system that achieves these objectives while addressing the issues and constraints noted in our task analysis.

The severe time constraints that naval forecasters work under are expected to increase due to current manning reductions and future regional forecasting (i.e., forecasts made for multiple remote sites). Using a monitoring paradigm, the MUM provides tools for quick assessment of model uncertainty. Forecasters can then decide the best use of their limited time. A forecaster wants to use, without correction, the model parameters which have performed well in the past and thus have limited uncertainty in the future. Other parameters may need further investigation through the use of personal knowledge to correct for uncertainty in a model forecast. Ideally, a forecaster should have automation tools that allow the selection of the best performing model (or ensemble member) plus the ability to modify model fields prior to forecast generation.

The MUM system is based on Java Server Page (JSP) and servlet technology. It is hosted at the Applied Physics Laboratory (APL-UW) on a Linux system running a Tomcat server. The model data used in the system comes from the UW SREF. This includes the global fields used for the SREF boundary conditions and the individual ensemble members of the SREF. This data is stored and archived on the APL-UW server.

Figure 2 outlines the data flow and software models that make up the MUM. Processing for model verification is routinely run in the background and maintained in a file that is called when the MUM JSP is downloaded to a user's browser. During the download of the interface page, a MUM model grabs the information that provides the color-coding of the stoplight graphics.

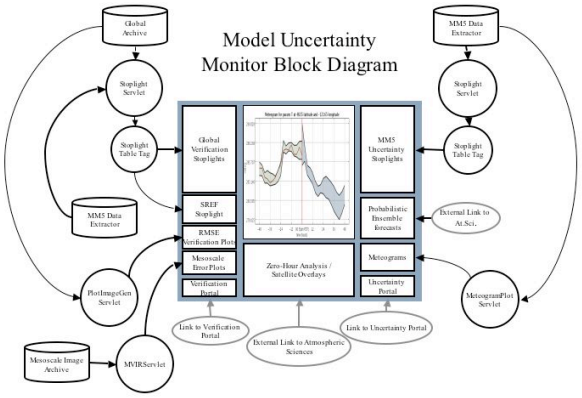

Figure 2. MUM architecture and data flow

5. MUM HUMAN-COMPUTER INTERACTION

While the MUM interface is still a work in progress, it currently produces model uncertainty assessments in real-time. The interface presents information in a past-present-future framework on the left, bottom center, and right control panels. As users select information links, a visualized representation is displayed in the center window.

Based on our task analysis, the forecasters' highest percentage of time was spent reviewing model initialization. Therefore, the default presentation is the current model initialization field overlaid on top of the most current satellite picture. The MUM currently pulls these products from the UW site: http://www.atmos.washington.edu/~bnewkirk/desc.html

Figure 3 shows the default view that is displayed inside a web browser after the selection of the MUM link: http://isis.apl.washington.edu/monitor/monitor.jsp. Clicking on the image can enlarge the default picture. In addition to the model field, the 24-hour wind forecast is shown compared to verification observations, along with the difference between forecast and observed pressure for that location. This assemblage of information provides the forecaster with the assessment of both the performance of the past model forecast and its initialization compared to the satellite picture. Further iterations of the MUM will include additional initialization assessment tools such as an interpolated model versus observation table for selected locations.

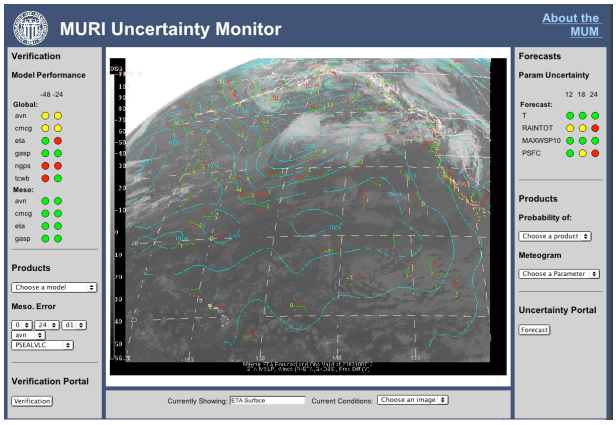

Figure 3. MUM with initialization information

On the left of the screen is model performance information. For the initial assessment, we use a stoplight paradigm, which is familiar to military personnel. These graphics provide pertinent information on the past performance of global models (upper section) and ensemble members (lower section). Models are judged based on their RMS error (Root Mean Squared Error) over a window of time in the past. RMSE is calculated by comparing a forecast against the corresponding zero-hour analysis. Comparing the most recent error result against the trend of error in the past generates the stoplight color. Green colored stoplights indicate low error, or good performance. Yellow indicates intermediate performance, and red indicates poor performance.

On the right side of the screen is uncertainty information about the future. As with the model performance section on the left, the top-level information is displayed with stoplight graphics, but these graphics are derived for a point (initially set at NAS Whidbey, KNUW). However, the model performance stoplights are derived over a geographic area (from global scale down to mesoscale).

Current research on ensembles has shown a relationship between the skill of the model prediction and the spread of the answers each member gives for a particular parameter. We use this spread relationship as a proxy for forecast uncertainty. The uncertainty stoplight table contains a number of stoplight graphics, the color of which attempts to classify the uncertainty of the model prediction. Green indicates less uncertainty or an indication of higher accuracy, while yellow indicates intermediate uncertainty. Red stoplights indicate high uncertainty for the corresponding combination of forecast hour and parameter. As the spread is a property

of the ensemble as a whole, uncertainty is not provided for individual ensemble members.

Ensemble spread directly corresponds to the stoplight color. If the ensemble spread is in the top 1/12th (91.66% and above) of all past spread values, the stoplight is red. If the spread lies between the top 2/12ths and 1/12th (83.33% to 91.66%) the stoplight is yellow. All other values are green.

The spread meteograms (shown in Figure 4) are new visualizations currently being tested in the MUM. These meteograms display information about MM5 ensemble performance for a single geographic location. Ensemble prediction data for a single parameter is shown over a time period of four days, where the most recent 00 hour prediction lies at the center, marked by a bright vertical line. The ensemble data is shown as a shaded region spanned by the minimum and maximum predicted value at the given time.

Data in the past 48 hours is a composite of three predictions; the -48 hour data is pulled from the 24-hour prediction initialized 72 hours previous, the -45 hour to -24 hour data (inclusive) is drawn from the prediction initialized 48 hours before the current 00 line, and the -21 hour to 0-hour values are from the ensemble forecast initialized at -24 hours. Future predictions are all from the forecast initialized at 00 hours, although the 0-hour value is not available for all parameters and may appear as a discontinuity.

In addition to ensemble spread, observation data appears in the left portion of the meteogram as a red-orange data plot. This is pulled from the METAR for the nearest observing station. In a well-tuned ensemble system this observation should typically fall within the range of the spread. If the observation routinely falls outside the spread, the forecaster can more easily pinpoint biases in the ensemble system.

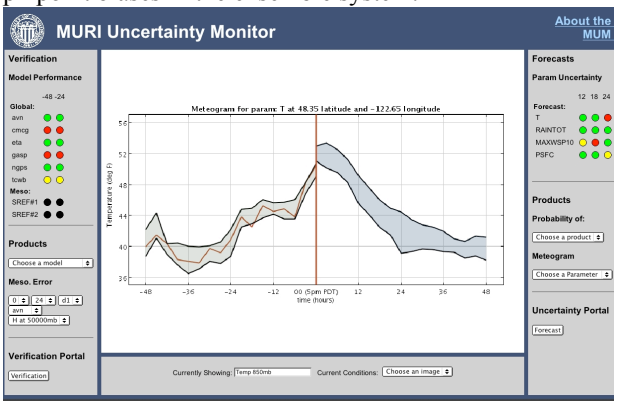

Figure 4. MUM with ensemble spread meteogram

Additional tools and visualizations will be added to the MUM. We are also investigating techniques for interacting with the probability distribution information that can be derived from an ensemble system. Navy operators may be interested in knowing the forecasted range of a particular parameter with a 99% certainty, or the value of the top two most probable forecast scenarios. The MUM will allow the user to interact with this type of probabilistic information.

6. CONCLUSION

We have developed a software framework and a prototype interface that assemble, process, and visualize uncertainty information for weather forecasters. This system is based on extensive observation and analysis of navy forecasters in their operational environment. The MUM will be used to test methods of presentation and user interactivity toward the goal of improving forecast quality, timeliness, and usefulness. Most importantly, we hope the MUM will encourage forecasters to use probabilistic information in new and innovative ways.

7. ACKNOWLEDGEMENT

This research is supported by the DOD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745. The support of the sponsor is gratefully appreciated.

8. REFERENCES

Edland A., Svenson O. (1993). Judgment and decision making under time pressure: Studies and findings. In *Time pressure and stress in human judgment and decision making,* Svenson O. and Maule A.M. (eds.) Plenum Press: New York; 27-40.

Jones, D. W., Ballas, J. Miyamoto, T. Tsiu, T., Trafton, G. & Kirschenbaum, S, (2002). *Human Systems Study on the Use of Meteorology and Oceanography Information in Support of the Naval Air Strike Mission*. APL-UW TM *8-02*. Seattle, WA.

Rogers & Monsell (1995). Costs of predictable switching between simple cognitive tasks. *Journal of Experimental Psychology: General* (124,2).

Grimit E & Mass, C (2002). Initial results of a mesoscale short-range ensemble forecast system over the Pacific Northwest. *Weather & Forecasting*. 17: 192-2005

Miller G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63: 81-97.

Trafton G.J., Kirschenbaum S.S., Tsui T.L., Miyamoto R.T., Ballas J.A., Raymond P.D. (2000). Turning pictures into numbers: extracting and generating information from complex visualizations. *International Journal of Human-Computer Studies* **53**: 827-850.