Evidence-based decision making as it applies to ensemble prediction system implementations

Some thoughts by Tom Hamill, ESRL/PSD, 4 May 2016
(updated 3 August 2016)

The UCAR review committee for the NCEP production suite, the UMAC (UCAR Model Advisory Committee) recommended in 2015, among other things, that: (a) NOAA develop strategic and implementation plans for weather and climate prediction, and (b) that NOAA model development follow an evidence-based decision making process. Our interest in this document is how this applies to ensemble prediction system (EPS) development.

We take as given that in the near future NOAA and EMC will expand upon the Next Generation Global Prediction System (NGGPS) Strategic and Implementation Plans to develop holistic Weather and Climate Modeling Strategic and Implementation Plans. These documents will outline the high-level thrusts of system development and how to make rapid progress consistent with Weather Ready Nation and NOAA Strategic Plans. Presumably ensemble experts will be consulted in the development of such a plan, and NOAA's ensemble prediction system (EPS) development will follow the guidance from such plans.

The other UMAC recommendation, an evidence-based process for decision making, is not yet in place for ensemble prediction systems. Definition of and use of such a process should help the NWS achieve better EPS products more quickly.

This document is thus offered for consideration by NWS and EMC management, a prospectus for how we could set up an evidence-based process for ensemble system implementations. What does "evidence-based decision making" mean in the context of EPS? Where does this decision making come into play in the development process? Below, we propose that there are typically four key decision points in the R2O process, points where a group of qualified individuals could be particularly helpful in making recommendations. Accordingly, this document thus suggests: (a) the potential composition of a review board; (b) the general criteria the review board should be considering when making recommendations, and (c) the decision points where their advice is needed.

1. Composition of a review board and their terms of operation.

The pre-eminent consideration for membership on the review board is a scientific understanding of ensemble prediction and its issues. While understanding of EMC's systems is desirable and at least one board member should be from EMC, we suggest casting a wider net to entrain the best subject-matter experts from other organizations, including NOAA/OAR, NASA, the US Navy, various universities, and perhaps experts from operational centers in other countries. We

suggest a panel of roughly five people, not so big as to be unwieldy but big enough to provide a diversity of experience. Following World Meteorological Organization practice, members would have 3-year terms which are renewable. Nominations for membership might be reviewed at a higher level, be it by the EMC Director, the NOAA Science Advisory Board, or other. This review panel would meet in person or via videoconference as needed, with their evaluation criteria and suggested decision points described below. The panel would provide written recommendations to the EMC Director and NOAA and NWS leaders.

2. Suggested criteria to evaluate a potential ensemble prediction system development.

(a) **Physically based**. Is the potential change to the EPS one that can be defended on scientific principles? Has the potential change been examined by relevant scientists, and have they agreed that the change makes the system more realistic?
(b) **Improvement**. Does the system with the potential change beat a previously agreed-upon baseline across a previously agreed upon number of metrics, in statistically significant amounts? Is it worth the expense of managing?
(c) **Code simplicity.** Does the change make the code simpler, or at least does it add complexity only in proportion to the increased physical resemblance to the real prediction system? Does the potential change make future (physically based) modifications easier? Is the code written in such a way that it can potentially be reused for other applications (regional models, climate models)? Does the code facilitate a potential reduction in the number of modeling systems that NCEP must run and maintain, allowing the model with the potential improvement to take up the product development from another modeling system? Is the software coded according to standards such that it can be readily adapted for operational use?
(d) **Code performance.** Does the potential change increase the CPU expense and/or disk storage, and if so, is the improvement in skill roughly concomitant with the increase in CPU and disk space?
(e) **Documented**. Is the methodology sufficiently documented so that it can be maintained?

3. Proposed stages and gates for the development process.

Here we assume that the development process is split into “stages” where work is performed, and “gates” where there is a critical review and a decision about whether to proceed on to the next stage. For a given system implementation, stages/gates 1-2 may be proceeding in parallel with multiple strands of research, and stages 3-4 are more typically with an integrated system that bundles improvements that have successfully passed through the first two stage/gates. Early stages may have a higher percentage of the work performed by OAR lab scientists or university investigators, later stages nearer to implementation are likely to have greater involvement from EMC.

Stage 1: Ideation. Based on many potential sources (development at other centers, conference results, research breakthroughs in academia, and/or the recent personal work of the

scientist or team), an idea is formed about how to improve the system. Presumably this idea addresses an agreed-upon need, such as the priorities outlined in the NGGPS implementation plan. The time allotted to this stage may vary significantly. The product generated at this stage is typically a research proposal, outlining the background research, the hypothesis, the proposed test plan with milestones and resources needed. The test plan is formulated with the first gate in mind, and the evaluation criteria discussed above. As the research is early on, the focus at this stage in on whether the proposed method is physically based and whether the literature provides supporting evidence of its potential.

> **Gate 1: Defense of proposal.** Whether this defense takes the form of peer review of a written proposal (e.g., with soft money) and/or an oral defense (e.g., for base funding) in front of a qualified panel, the expectation is that there is a decision point where higher management, perhaps informed by a panel of experts, either approves the project for the next stage, or not. Approval may be contingent; the researcher(s) may be asked to provide more evidence, to submit a revised proposal that trims the scope of the project, or that modifies the research plan according to guidance from the panel.

Stage 2: Preliminary experimentation. The researchers now execute their proposed test plan, developing the new improvement, testing it against an agreed-upon baseline. Likely the researchers will engage in a substantial iterative process, where they learn about deficiencies, de-bug code and/or modify the hypothesis and the technical approach. During this phase, the expectation is that the scientist will be regularly consulting with experts and peers, subjecting their intermediate results and code to scrutiny, accepting and acting on relevant feedback. The product generated from this stage will be a set of preliminary results and a report/presentation that addresses the evaluation criteria above.

> **Gate 2: Decision on whether preliminary experimentation warrants pre-operational development.** The scientist or team involved in the research presents their preliminary findings to a review panel. This panel, with strong representation from the operational center (EMC) and other relevant parties (STI, NCO, CPC, other relevant NCEP centers), will evaluate the results and make the decision as to whether the results presented show enough promise to proceed with operational development. The review panel may make suggestions for this operational development phase; for example, they may recommend testing of the system at a higher resolution, or testing in conjunction with other related developments. The primary focus at this stage is whether the preliminary experimentation provided an improvement, with some attention to code performance and code simplicity.

Stage 3: Pre-operational development. Presuming a positive recommendation at gate 2, in this stage the code developed in stage 2 is adapted to the operational environment and tested more rigorously. This testing might include experiments at the anticipated operational resolution, testing over a broader set of cases, testing over a broader set of metrics, and/or against more stringent reference standards (perhaps against the anticipated next version of the

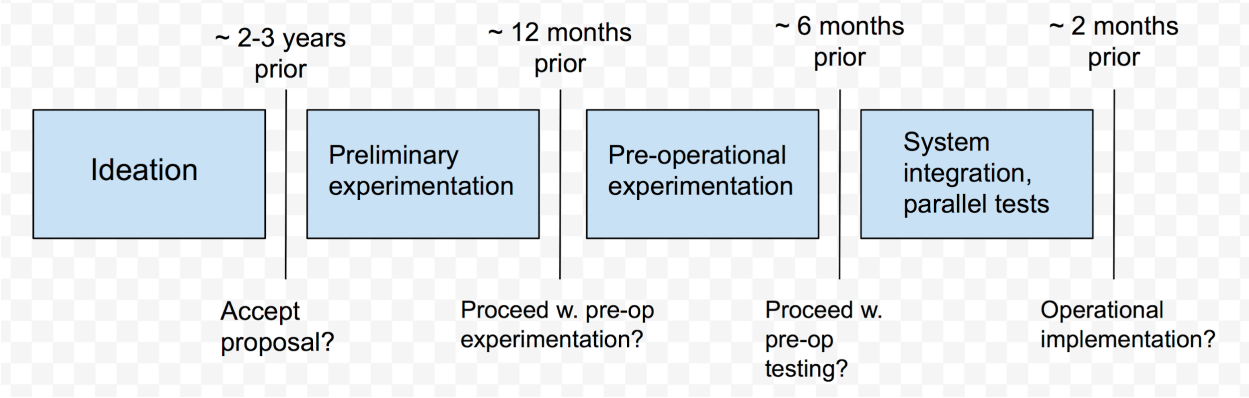
system, with other improvements). The scientists involved will prepare documentation for the next formal review, addressing the evaluation criteria above.

> **Gate 3: Decision on whether to proceed with pre-operational testing**. NWS management and chosen outside experts will evaluate the review material, making a recommendation as to whether to proceed. Other options may be discussed (return to the previous stage for more development, delay one implementation cycle, etc.). The focus is on performance, code simplicity and performance, and adequacy of documentation.

Stage 4: System integration and parallel testing. EMC staff will now take the software and documentation previously prepared in stage 3, adapting it and merging it with other proposed software enhancements that have also proceeded through gate 3. They will develop an integrated new version of the system, presumably test the integrated version to make sure all components are working properly, and validate the system performance over a moderate number of cases relative to an established reference standard. Presuming these results are satisfactory, the upgraded software is now sent to NCO for their final adaptation and is made into a formal "parallel" model run, and is compared side-by-side with the current operational model over the chosen test period. Results are synthesized into a report.

> **Gate 4: Implementation decision**. NWS Senior management and chosen advisors make a decision as to whether to proceed with the operational implementation. If they do, responsibility for the final steps becomes more of a responsibility of NCO.

A rough notional GANTT-type chart for stages and gates is provided below. The length and number of stages and gates may differ with the scale of the proposed project.



Suggested next step:

We suggest consideration of this by EMC, NCEP, and STI leadership. Assuming that NCEP is in agreement with the fundamental details here, we suggest that EMC, ESRL, and other

interested parties work together to: (a) more concretely define the concept of operations; (b) draft a terms of operation for the review panel; (c) define evaluation criteria, and (d) set the initial composition of review board.