21st International Congress on Modelling and Simulation, Gold Coast, Australia, 29 Nov to 4 Dec 2015
www.mssanz.org.au/modsim2015

Thoughts on spatio-temporal uncertainty metrics motivated by input sensitivity in the Spark bushfire spread model

C. Huston [a], **C. Miller** [a], **J. Hilton** [a], **F. Woolard** [a] **and E. Campbell** [a]

[a]*Digital Productivity Flagship CSIRO, Clayton, Victoria, 3169*
Email: carolyn.huston@csiro.au

Abstract: Bushfires are highly complex events to both measure and predict. We want our models to predict certain observable features of fires, but which features to use is a question with many possible answers. It is important to understand both the skill of our bushfire spread prediction models, and the level of uncertainty inherent in their results. However, this can be difficult to quantify even when the desired features are known. One challenge is obtaining complete observed data. A second is knowing how to determine appropriate comparative metrics for predictions and observations that evolve over space and time. Further to this, we need to consider the intended use case of the metrics. Viability of a metric may depend on whether it is needed to inform: machine decision-making, researchers familiar with the subject area, or public stakeholders and decision makers. Here, we illustrate some problems and solutions that we have encountered in determining metrics appropriate to bushfire modelling, along with our proposed approach to measuring model skill. In particular we focus on metrics evaluating predicted bushfire perimeters, as opposed to other aspects of fire behaviour such as fire intensity, flame depth and height, etc.

We model the evolution of a bushfire perimeter through time, given sparse data and incompletely observed perimeters. Considering the intended use cases of our comparison metrics, we have to communicate our model skill with human stakeholders and also use these metrics to improve our models in a relatively automated way. This adds an extra layer of complexity in selecting and applying metrics. Rather than focusing on a single metric that will likely never meet all our needs, we instead propose a standard approach to the development of metric sets appropriate to a problem. Our work utilises ideas from goodness-of-fit testing in the context of posterior and prior predictive approaches. We use this approach to simultaneously develop metrics for both human stakeholder and computational model development purposes.

In this paper, we demonstrate both visual and computational two-dimensional metric solutions, as well as a scalar metric appropriate for different computational purposes. The scalar metric is one originally proposed in Baddeley (1992) which has been previously applied to weather forecasting (Gilleland et al., 2008). We also provide details on our standard approach, which could be extended to other spatio-temporal and complex models.

Keywords: *Model skill, goodness-of-fit, scientific communication, spatio-temporal, bushfire modelling*

1 INTRODUCTION

Assessing, understanding and communicating model skill results is an extremely important, although sometimes ignored, component of scientific research. Without model skill assessment, there is no opportunity to evaluate and build trust in the modelling results.

Many commonly reported and considered metrics of model skill such as the coefficient of variation (R^2), χ^2 goodness-of-fit test, or Root Mean Square Error (RMSE) evolved in the early and mid portions of the twentieth century (D'Agostino, 1986). Such classical ideas often relate to hypothesis tests and p-values and represent an overall, or summary, measure of discrepancy between models and truth/observations. The choice of such broad tests across the results of a model has a two-fold motivation. First is that it is often useful to have an overall picture of model performance, both as an assessment of the model and a tool to compare between the performance of alternate models. Second, at the time they were developed they represented an efficient use of available computing resources; generating multiple measures able to consider many different failure modes would have been time-consuming with the existing computational resources.

Computation has now become less expensive allowing the development of more complex simulation models and, consequently, so too has our ability to explore the skill of these models. In Bayesian statistics in particular this has led to a re-examination of the purposes of model goodness-of-fit, and a proliferation of ideas and approaches regarding how it can be achieved. In Gelman et al. (2003) the authors explore some of the reasoning behind goodness-of-fit and model checking, identifying that the key question is rarely 'Is our model true or false?', which is the answer given by many broad model skill statistics. Instead 'Do the model's deficiencies have a notable effect on the substantive inferences?' is found to be a more central question. The authors then advocate for the use of posterior predictive approaches to model fit, where the core idea is that data simulated based on a model should be representative of observed data. Multiple simulated data points should contain any observed data points. While traditional goodness-of-fit test statistics can be calculated based on such posterior predicted simulated data, it is equally possible to identify highly specific test statistics to evaluate a model's effectiveness at answering questions of importance.

While the textbook focuses on numerically defined *test statistics*, in Gelman (2003), the authors extend the idea of model checking to include the role of, and development of tools for visual model checking. This in turn aligns to ideas about plotting related to exploratory data analysis where scientific learning is considered an iterative process between criticism and estimation (Box, 1980). As enablers of this process, visual representations of data are being recognized as a type of metric themselves, useful in conveying model skill.

With a spectrum of choice extending from scalars to detailed visualizations, it is challenging to identify a single best approach to reporting on model skill. If one chooses the scalar approach, it might have the advantage of being recognizable to colleagues and simple to make judgements about. Communicating the deeper meaning of such a number to outside stakeholders in a persuasive way is often challenging in this case. Conversely, assessing a visualization, or comparing visualizations might be accessible and engaging to non-specialist stakeholders. Such a visualization also might not easily lend itself to a determination of whether one model formulation is objectively better than another.

Here, we advocate for a unified *process* for identifying appropriate tools for identifying model skill. Rather than advocating for a best metric to be used in all cases, we propose that identifying a fit-for-purpose metric through our process will give best results.

We feel a particular benefit of such an approach is that it opens up the use of visual tools in model skill assessment. The revolution in new technologies such as personal devices, HD pictures and video; computing power; etc. has resulted in much of society relying on more immersive experiences to engage them. If we can engage people in scientific debate and thinking through use of similar tools, it serves the purpose of broadening the audience of scientific model outputs, and strengthening the role of science in society.

2 CASE STUDY: BUSHFIRES

Being able to accurately predict locations of a fire front has many potential benefits such as better information to create evacuation plans, the ability to support suppression strategies, and obtaining scientific insight about bushfire behaviour. One of the challenges in assessing predictions of bushfire behaviours is incomplete data of the fire propagation over time. For example, fire perimeter observations might only be taken at discrete/point locations, such as when an observer directly reports a fire reaching a location. Even satellite and other images of a fire can be obscured by smoke, making it difficult to observe a complete perimeter. Additionally, satellite

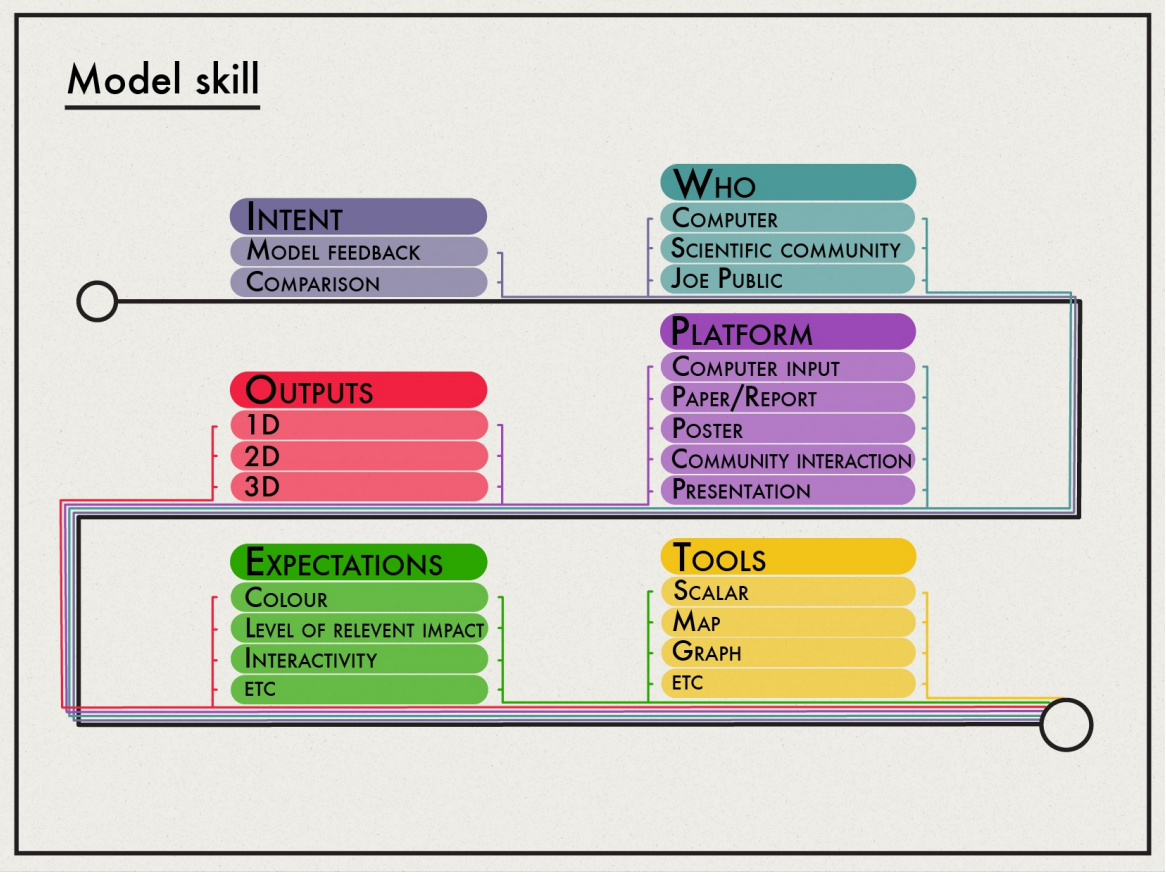

Figure 1. A process for identifying features for appropriate measures of model skill.

or other scanning images are often only available at sparse time-points during the progress of a fire.

We demonstrate some concepts around model skill based on outputs from a level-set bushfire solver, Spark (Miller et al., 2015). The model used in Spark to propagate a fire front relies on inputs including a model for a fire front rate of spread and contributing factors such as fuel type, topography, etc. The model output is a fire perimeter that evolves through time. We consider model skill of the solver from both a computational research perspective and from a scientific communications perspective.

The goal is to integrate the Spark modelling engine into a fully Bayesian hierarchical model where observed fire perimeters are used to provide inference about modelling parameters, such as wind, and consequently to better predict fire spread. From a modelling perspective, factors like scale (size), translation (left-right shift), and rotation (angular offset of spread) were all important features that needed to be captured in a scalar measure of model skill. A number of scalar metrics for model skill in the fire modelling domain have been proposed and demonstrated (Filippi et al., 2014). Other methods for shape analysis appropriate to fire perimeters also exist in the analysis of digital images (eg. Baddeley (1992)) and statistical shape analysis in morphology (Dryden and Mardia, 1998).

More subtle and nebulous features of fires also have to be considered, especially in the eyes of fire experts. An end-goal for the Spark solver and associated Bayesian method is to support decision making that supports resilience to bushfires. Consequently, fire experts, local communities, etc. need to be able to understand model outputs well-enough to either trust them, or provide us with critique about where the model lacks skill.

3 Methods

In Figure 1 we illustrate path of decisions that can be made around selecting an appropriate method of measuring and reporting model skill. We identify a number of decision points on the pathway, and include non-exhaustive lists of factors to take into consideration at each of these points. It is assumed that once a choice

has been identified at a decision point that it carries through and influences available choices in the remaining decisions. The goal of following the process is, at a minimum, to identify key features that are needed for a model skill approach to be fit for purpose. Jakeman et al. (2006) recommend the consideration of a range of indicators of model effectiveness, and provide a list of possible indicators for consideration. More aspirationally, our goal would be to demonstrate or identify actual methods that could be used, along with information on the strengths of each to further simplify the selection process. Each of the steps on the pathway is outlined in greater detail below.

3.1 Intent

As discussed in Section 1, all model skill methods relate to an overall goal of model checking. Despite this, the focus of this model checking can vary in different circumstances, and we identify sub-goals as including direct comparison between models and facilitating critique and feedback about a model. Comparison between models can be further separated into considerations such as "Am I doing better with these operational settings?" to "Am I doing better than competing models?"

3.2 Who

At this decision point, the final consumer of the model skill representation needs to be identified. Although we call this the 'who' decision point, we recognize that the final consumer of a model skill metric may not be human. For example, the end recipient might be a computer program performing automated decision support tasks. Even when the intended recipients are human, consideration needs to be taken in identifying different sub-groups of individuals. For example, scientists researching in the same area will bring different background knowledge about the research area, common problems, status quo approaches, etc., than a member of the general public would. To use the same model skill approach for both specialists and the general public might require an investment in teaching the general public core science/skill ideas around the research area; alternately, it might indicate that an alternate representation of model skill would be appropriate.

3.3 Outputs

Identifying exactly what a model produces as output, or what it can be made to produce, is important in determining what model skill approaches will be appropriate. A regression model that produces coefficient weightings for variables is a substantially different output to a shape, such as is generated by our bushfire model. We particularly highlight the importance of the dimensionality of the output, as when we reach sub-Section 3.6 what tools are available to report on model skill will depend on what outputs can be created (or imagined).

3.4 Platform

Where model skill will be presented can place expectations on what reporting tools are appropriate. Even within a specific platform, such as a scientific conference, presentation type diverges between posters and oral presentations. Answering *who* in sub-Section 3.2 also helps inform which platforms are appropriate.

3.5 Expectations

Knowing the choices made at the above decision points provide insights about potential audience expectations. An audience of the general public might hope for an immersive experience such as an engaging illustration or a video; if a model outputs a scalar such model skill options will be impossible. Conversely, even if an output is spatio-temporal and can be animated, if the platform is a poster session at a scientific conference a static approach must be used.

3.6 Tools

Knowing available tools in terms of existing model skill approaches is important. However, the best solution might be to develop a new approach appropriate to all previous decisions. Scalar approaches such as R^2 and RMSE as were mentioned in Section 1 are common and generally well-understood by appropriate scientific audiences. They do have limitations though, even for this purpose. Models are becoming increasingly complex as we better understand how to use improvements in computational power, so results that can be reported in space-time dimensions are increasingly common. While more traditional approaches can sometimes be applied, there is much more scope with such results to use focused metrics to understand where models are

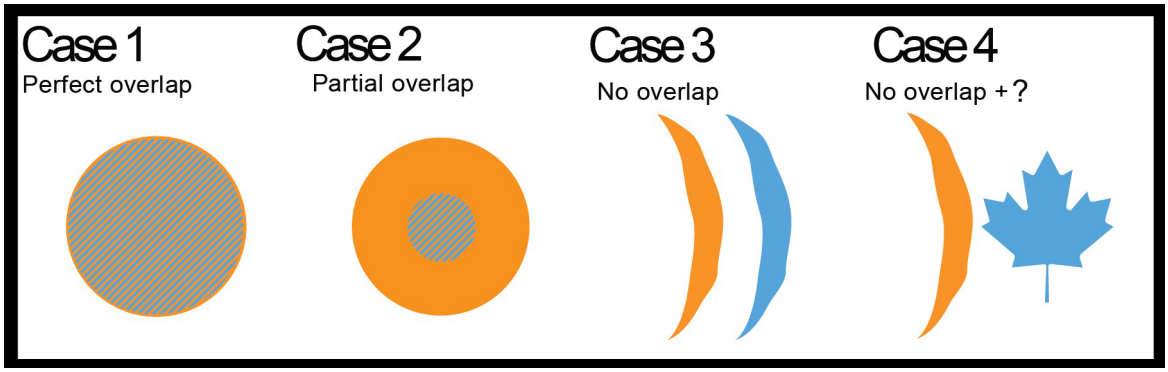

Figure 2. Generalized outcomes that a scalar model skill assessment has to differentiate between.

weakest. Alternately, new tools to visualise data involving maps or animations can be used.

4 Application to Bushfire Case Study

To demonstrate the process discussed in Section 3 we consider two different use cases for model skill based on the bushfire modelling case study introduced in Section 2. In the first case we are looking for a measure of model skill that can be used to automatically update bushfire modelling parameters by assessing which parameters lead to better or worse results compared to the truth. In the second case, we are soliciting general feedback from fire domain specialists about how a model 'looks' relative to their expectations about fire behaviour.

4.1 Computer appropriate approach

As mentioned in Section 2 our goal is to find a single metric appropriate to use as part of a Bayesian hierarchical model. Such a model is based on creating multiple simulations. Prior distributions would be elicited for input parameters such as wind strength and direction, fuel, etc. based on some combination of empirical data and expert judgement. Parameter sets drawn from such priors would then be used to identify those most consistent with the observed data as assessed by an appropriate fit metric.

When considering the decision points it was *intent*, *who*, *outputs*, and *expectations* that were the biggest differentiators between model skill methods we considered. Our goal was comparison between model runs in order to determine which combinations of model parameter inputs lead to results that are closer to observational ones. As our *who* was a computer, we needed to find a scalar representation of fit which would be easy to assess algorithmically. Despite the goal for a scalar measure of model skill, the actual outputs of the Spark model are fire perimeters, two-dimensional shapes, which are then evolved through the time dimension. This high dimensional information had to be summarized.

In terms of our *expectations* of the model skill metric, it had to fulfil a number of key criteria. As stated, it needed to summarize multi-dimensional information into a scalar. We also wanted it to be able to differentiate between certain key potential cases, which are summarized in Figure 2.

Somewhat equivalently, we want our metric to be a true distance measure such that $d(A,B) = d(B,A)$; $d(A,B) > 0$ if $A \neq B$; $d(A,B) = 0$ if $A = B$; and $d(A,B) \leq (A,C) + d(C,B)$ where A and B are shapes to be compared, and C is an intermediate shape between the two.

Finally, the metric needed to be computationally efficient, and account for missing data occurring in the observed fire perimeters.

We considered a number of area based metrics (Filippi et al., 2014); statistical shape analysis approaches (Dryden and Mardia, 1998); and one approach developed for black and white image analysis (Baddeley, 1992). Ultimately, due to the clear criteria that had emerged in our development process, the Baddeley Δ metric was selected as fit-for-purpose as described below.

$$\Delta^p(A,B) = \left[\frac{1}{n(X)}\sum_{x\in X}|d(x,A) - d(x,B)|^p\right]^{1/p} \quad \text{where } d(x,A) = \min_i(x,A_i) \text{ and } A_i \in A$$

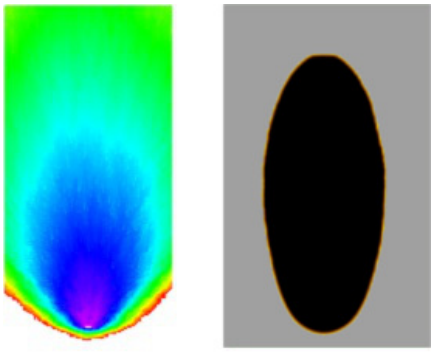

Figure 3. Two fire visualizations where the audience implicitly compares model output results to known fire behaviour. In the figure on the left, colour represents the length of time it takes for fire to burn a region in the picture. In the figure on the right we show burned and unburned areas at a timestep in the model.

Here Δ^p is considered up to a fixed constant or cut-off distance $c > 0$ and where A represents the set of points on shape A (or B), and x is a point in the set of points X representing the area surrounding A and B up to the cut-off distance c. Following this, $n(x)$ is the number of points assessed in X. The constant p is a power that the metric can be raised to. For example, $p = 2$ yields an equation similar to the standard difference between points. In a traditional distance context p is modified to manage the influence of skewness or outlying observations.

Advantages of the Δ metric over others are that the calculation yields a scalar (statistical shape/Procrustes methods did not). It is also a proper distance metric (which most area based metrics are not because they break down when shapes don't overlap). It is fast to compute as the $d(x, A) = \min_i(x, A_i)$ and $A_i \in A$ are calculated as part of the narrow-band algorithm used in the Spark solver. Finally, because points in a shape are referenced to points $x \in X$, there is not requirement that shapes being compared have equal numbers of points to compare, or that they have a continuous representation. This is useful in the context of missing data which is common with observed fire fronts.

4.2 Domain specialist appropriate model skill representations

Contrasting our computer appropriate metric, the intent of demonstrating model skill to domain specialists is to solicit model feedback. Acting on this model feedback will hopefully build confidence in model results so that it is adopted by practitioners. In this case we are presenting model skill to the fire science community, not the mathematical modelling community.

For this, the spatio-temporal dimensionality of the data can be leveraged to make model outputs reflective of what people expect to see when they look at a picture of fire. The platform chosen should be one that allows for interaction between the audience and the presenting scientist such as through a presentation, or at a formal poster session. To facilitate feedback in a timely way, visual representations of the data that are familiar to the audience should be used. Colour selection plays a part, especially when certain colours will be interpreted as 'hotter' than others. Maps or fire perimeter contours would be a traditional way of recording observed fire information, so our outputs should be relatable to this. Trying to explain how to interpret scalar Δ metric results would be less timely and likely less approachable.

In Figure 3 we showed some test images that are similar to ones that resulted in very valuable learnings. In the figure on the right, we see a fire perimeter that looks like a blob. Our audience was not very convinced that our model produced results that looked like 'fire.' From further conversations based on related images, managing fire curvature (how rounded or sharp the perimeter can be) resulted in a more acceptable image. In the image on the left, parameter inputs are varying rather than constant. Particularly in the blue area of the fire a feature called 'fingering' is visible. As mathematical modellers we were unaware that this was a desirable feature to have captured until an expert was able to relate the image to real fires they had seen. In terms of obtaining feedback from a scientific development of the model standpoint, this was a very useful conversation.

5 CONCLUSIONS

Here we demonstrated two different use cases where representations of model skill were desirable. The first use case required a measure for the shape of a bushfire perimeter relative to an incompletely observed 'true' fire perimeter. The second was to communicate model skill of our bushfire model results to a broader audience in order to solicit feedback regarding perceived strengths and weaknesses of the modelling approach. In one case we created a scalar metric appropriate to the use goal, but which is difficult to interpret intuitively or in isolation of related measures. In the other, where we wanted information from experts' intuitions, we chose a visual discrepancy approach.

For simplicity it is attractive to imagine a parsimonious approach where a single measure of skill is sufficient to assess and validate a model, and also can be effectively communicated to diverse audiences. After considering how to attempt this, our solution was not to advocate for a single *measurement* to assess model skill; it is to advocate and trial a single *process* on how to illustrate model skill for diverse audiences and purposes.

Engaging in such a process to create use specific products requires additional thought and time relative to a catch-all approach. As a first attempt, refinement is still required to flesh out our process approach. Despite this, we feel that a process approach like we demonstrated will ultimately yield more desirable outcomes to situations where measuring model skill and reporting are required.

REFERENCES

Baddeley, A. (1992). Errors in binary images and an lp version of the hausdorff metric. *10*, 157–183.

Box, G. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A 143*(4), 383–430.

D'Agostino, R. B. (1986). *Goodness-of-fit-techniques*, Volume 68. CRC press.

Dryden, I. L. and K. V. Mardia (1998). *Statistical shape analysis*, Volume 4. Wiley Chichester.

Filippi, J.-B., V. Mallet, and B. Nader (2014). Representation and evaluation of wildfire propagation simulations. *International journal of wildland fire 23*(1), 46–57.

Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review 71*(2), 369–382.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (2003). *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton.

Gilleland, E., T. C. Lee, J. Halley Gotway, R. Bullock, and B. G. Brown (2008). Computationally efficient spatial forecast verification using baddeley's delta image metric. *Monthly Weather Review 136*(5), 1747–1757.

Jakeman, A., R. Letcher, and J. Norton (2006). Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software 21*(502), 602–614.

Miller, C., J. Hilton, A. Sullivan, and M. Prakash (2015). Spark–a bushfire spread prediction tool. In *Environmental Software Systems. Infrastructures, Services and Applications*, pp. 262–271. Springer.