Environmental Modelling & Software 83 (2016) 276–285

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Indices for the evaluation of wildfire spread simulations using contemporaneous predictions and observations of burnt area

Thomas J. Duff*, Derek M. Chong, Kevin G. Tolhurst

School of Ecosystem and Forest Sciences, Faculty of Science, The University of Melbourne, Burnley 3121, Australia

A R T I C L E I N F O

Article history:
Received 19 March 2015
Received in revised form
4 May 2016
Accepted 4 May 2016

Keywords:
Area Difference Index
Bushfire
Fire spread model
PHOENIX RapidFire
Jaccard's coefficient
Kappa
Precision
Recall
Shape Deviation Index
Sørensen similarity
Spatial validation
Verification
Wildfire

A B S T R A C T

Methods to objectively evaluate performance are critical for model development. In contrast to recent advances in wildfire simulation, there has been limited attention to evaluating fire model performance. Information to validate fire models is typically limited, commonly to a few perimeter observations at a small number of points in time. We review metrics for comparing two burnt areas at a point in time: observed and predicted. These are compared in an idealised landscape and with a case study evaluating the performance of simulations of an Australian wildfire. We assessed: Shape Deviation Index (SDI), Jaccard's coefficient, F1, Sørensen's Similarity and Area Difference Index (ADI). For decomposing fit into error components (overprediction and underprediction) we assessed the partial indices of SDI and ADI, Precision and Recall. The various metrics were evaluated for their ability to represent error and their suitability for use in model improvement frameworks.

1. Introduction

Forest fires are a frequent event in parts of the world that have hot, dry summers and cool wet winters. While the vegetation in these areas is adapted to periodic disturbance, these fires regularly cause losses of human life and property. The economic impacts of catastrophic wildfires, such as the recent 'megafires' in Australia (Cameron et al., 2009) and Greece (Pyne, 2008), can be substantial. The Australian black Saturday fires of 2009 resulted in 173 deaths and an estimated A$4 billion in damages (Teague et al., 2010). As a consequence, land managers have been investing in technologies that can provide a better picture of fire behaviour and subsequent impacts. One such growing field has been the development of dynamic fire spread models; models designed to predict the spread of fire across the landscape (Sullivan, 2009). A number of single dimension 'forward rate of spread' (FROS) fire models were developed throughout the 20th century, empirically fit against experimental fires. These were extended into spatially and temporally explicit fire simulators such as FARSITE (Finney, 2004), PHOENIX RapidFire (Tolhurst et al., 2008) and Prometheus (Tymstra et al., 2010). While these are being used by land managers to predict the spread of wildfires, such fires are generally outside the conditions under which the models were developed. Consequently, they should be conditionally verified if their results are to be relied upon (Jakeman et al., 2006).

A key part of any verification is the selection of basic performance criteria (Bennett et al., 2013). FROS models are amenable to this as they provide a single prediction that can be verified against observations using direct value comparison (Bennett et al., 2013). However, when FROS models are incorporated into simulation models the outputs are complex as they represent an event driven by temporally changing influences through heterogeneous landscapes. While the FROS models from which the simulation models have been derived may have been experimentally verified, this does not mean that the outputs of the derivative simulation models

* Corresponding author.
E-mail address: tjduff@unimelb.edu.au (T.J. Duff).

http://dx.doi.org/10.1016/j.envsoft.2016.05.005

are necessarily valid – particularly as the simulation models must predict the effects of changing conditions (in time and space) and additional fire behaviour to forward spread (such as flanking spread).

Additionally, as the simulation models are intended for predicting wildfires, they are almost invariably being applied under hot and windy weather conditions that are outside the experimental conditions under which their algorithms were developed. For managers to understand the effectiveness of simulation models for predicting wildfires, they need to be conditionally verified for the conditions under which they are intended to be used. Assessing model performance against real events is an important part of addressing a core question of model evaluation '*Does its behaviour approximate well that observed in respect of the real thing?*' (Parker et al., 2002).

One of the issues with evaluating the performance of models for simulating wildfires is that such fires are rapid, transient events that occur with little notice. When they occur in populated areas managers are typically focussed on fire suppression, human safety and asset protection. Consequently, there is usually limited information describing the spread of fire through time. Often, all that is available for model verification is the final burnt area (Duff et al., 2014). While there may be limited information available for any individual fire, if data from enough fires are obtained, conclusions can be made about model performance. To be able to do this, methods of the assessing model performance solely from predicted and observed burnt areas are necessary.

While there has been rapid development in fire spread simulators, there has been limited attention to metrics for evaluating model performance. In practice, most assessments have been limited to broad subjective descriptions that cannot be easily independently verified (Berjak and Hearne, 2002; de Vasconcelos et al., 2002; Stratton, 2006; Johnston et al., 2008). Of the studies into quantitative evaluation that have been done, the primary focus has been on the ability of the models to correctly predict the area burned in a fire by comparing contemporaneous predicted and observed areas (Fujioka, 2002; Arca et al., 2005; Cui and Perera, 2010; Duff et al., 2012; Filippi et al., 2014; Kelso et al., 2015). Despite this commonality in focus, there are no indices of performance that are being consistently used.

To investigate potential indices for assessing overall model performance (i.e. evaluating correspondence between simulated and burnt areas), we compared a number of indices using a 'perfect world' approach and a real-world case-study. These are Jaccard's coefficient, Sørensen's similarity, the Shape Deviation Index (SDI), the Area Difference Index (ADI) and F1. Additionally, we evaluated the partial (i.e. those that can discriminate overpredicted area and underpredicted area) metrics of SDI (SDI_{over} and SDI_{under}) and ADI (ADI_{over} and ADI_{under}) as well as Precision (a measure of overprediction) and Recall (a measure of underprediction). The indices are described in more detail in the following section. The assessments were done in the context of their suitability for assessing existing models (rather than designing models) and consider the need for indices that are robust, unbiased, and efficient and provide acceptable discrimination of predictive performance (Jakeman et al., 2006). Needs specific for fire analysis include that the indices provide for the objective comparison of two fire affected areas for the same point in time (predicted and observed), be robust in the face of a wide range of fire shapes, locations and spread patterns, and be suitable for use in systematic improvement processes. We review the candidate indices for model evaluation and provide some recommendations based on our findings.

2. Methods

2.1. Technical background

Overall, single value performance metrics are valuable for the rapid evaluation of predictions against observed burnt areas. While they don't provide a high level of detail on the sources of error, they can be valuable for verifying and comparing models by using a number of case studies (Duff et al., 2012). We focus on single metrics designed for comparing two burnt areas (intended to represent a prediction and equivalent observation) at a point in time. We do not evaluate temporally dynamic metrics – while these are invaluable for model improvement, they require substantially more information.

For such evaluations, the fire affected (burnt) area is typically the property of interest when comparing simulated fires (S) against reference 'real' fires (F). It can be delineated spatially by closed polygons indicating the perimeters of burnt area. When comparing simulated and observed fire perimeters, there are three key properties that are recognised; the intersection area (I; the area common to both fires that would be considered correctly predicted), the overpredicted area (OE; the area predicted to be impacted by S that was not actually burned by F) and the underpredicted area (UE; The area burnt by F but not encompassed in S). These are presented schematically in Fig. 1.

Filippi et al. (2014) suggested a number of indices that are potentially suitable for evaluations, including Jaccard's coefficient (Jaccard, 1901) and Sørensen's similarity index (Sørensen, 1948). Jaccard's coefficient (1) was designed as a metric for comparing similarity between floral communities and is calculated as the ratio of the intersection area (I) and the union of the fire shapes being evaluated at time t (Jaccard, 1901).

$$
\begin{aligned}
Jaccard's\ coefficient(t) &= \frac{I(t)}{(F(t)+S(t))-I\,(t))} \\
&= \frac{I(t)}{(I(t)+UE\,(t)+I(t)+OE(t))-I(t))}
\end{aligned} \tag{1}
$$

Sørensen's similarity (2) is very similarly structured to Jaccard's coefficient, being a ratio of the intersection area and the sum of the shapes being evaluated (Sørensen, 1948). As with Jaccard's coefficient, it was designed for floristic analysis.

$$
\begin{aligned}
Sørensen's\ similarity &= \frac{2\,I(t)}{(F(t)+S(t))} \\
&= \frac{2\ \ I(t)}{(I(t)+UE(t)+I(t)+OE(t))}
\end{aligned} \tag{2}
$$

Also suggested was the Kappa coefficient (Arca et al., 2005), the

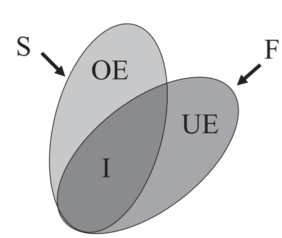

Fig. 1. Schematic representation of the relationships between a reference fire area (F) and a simulated fire area (S). The features indicated are the intersection area of the fires being compared (I), the area Fire metrics evaluated using SDI and ADI. F = Test fire, S = Reference fire, OE = Overestimate, UE = Underestimate, I = Intersection. Reproduced from Cui and Perera (2010).