# MACHINE LEARNING BASED SURROGATE MODELLING AND PARAMETER IDENTIFICATION FOR WILDFIRE FORECASTING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Simulating wildfire propagation in near real-time is difficult due to the high computational cost and inappropriate choices of physics parameters used in the forecasting models. In this work, we first proposed a data-model integration scheme for fire progression forecasting, that combines deep learning models: reduced-order modelling, recurrent neural network (Long-Short-Term Memory) and data assimilation techniques. Capable of integrating real-time satellite observations, the deep learning-based surrogate model run about 1000 times faster than the Cellular Automata model used to forecast wildfires in real world scenarios. We then addressed the bottleneck of efficient physics parameter estimation by developing a novel inverse approach, relying on data assimilation in a reduced order space. Both the fire prediction and the parameter estimation approaches are tested over recent massive wildfire events in California with satellite observation from MODIS and VIIRS to adjust the fire forecasting in near real-time.

.

## 1 INTRODUCTION AND RELATED WORK

There is a significant increase of wildfire frequency world-widely over the past decades, causing losses of lives, huge economic cost and lethal effects of air pollution (Chen et al. (2021)). Running physics-based simulations for large fire events is computationally expensive due to the complexities of physical models and geographical features. A key challenge of wildfire forecasting is to develop surrogate models, capable of i) efficient near real-time fire forecasting, and ii) incorporating real-time observations for system updating/adaptation. Ideas of using machine learning approaches, namely recurrent neural networks (RNN) in the reduced order space for predicting high-dimensional dynamical systems have been employed in a wide range of engineering problems such as computational fluid mechanics (CFD) (San et al. (2019)) and air pollution estimation (Casas et al. (2020)). In this work, we propose a modular approach, combining reduced-order modelling (ROM) (e.g., Principal Component Analysis (PCA) and Convolutional autoencoder (CAE)), recurrent neural network (RNN), data assimilation (DA) and error covariance tuning for forecasting the burned area of large fire events. In this paper, DA is performed in the ROM space. In particular, we use Latent Assimilation (LA) which is a DA method performed in the latent space of an Autoencoder (AE) (Amendola et al. (2020)). This approach is tested in two recent large fire events in California where a CA fire simulator is employed to generate the training dataset for the reduced-order surrogate modelling and satellite observations from Moderate Resolution Imaging Spectroradiometer (MODIS) and Visible Infrared Imaging Radiometer Suite (VIIRS) (Giglio et al. (2016)) are used as observation data in the DA. However, a limitation of this model is that it strongly relies on the physics-based simulations which are often not sufficiently accurate due to an inappropriate choice of physics parameters. In fact, another important challenge of the state-of-the-art fire modellings (both raster- and vector-based approaches) is the model parameter estimation, which is often obtained via numerical experiments or the resolution of a fire/region specific optimization problem (e.g.,Alexandridis et al. (2008); Ambroz et al. (2018)). The requirement of parameter/hyperparameter tuning for different fire events (often "manually" as described in Lautenberger (2013)) is not only computationally demanding but also poses a limitation in terms of generalizability. On the other hand, applying directly

inverse methods such as DA or Approximate Bayesian Computation (Vadera et al. (2020)) for wildfire problems can be extremely difficult, due to the high-dimensionality/complexity of the physics- or ML-based forward models. In this work, we introduce a novel LA method to tackle the bottleneck of parameter estimation. Our method, called Generalised Latent Assimilation (GLA) is an extension of a LA method, and thanks to a local smooth approximation function, it is capable of combining information sources (namely background vector and observations) issued from different latent spaces. In this paper, we show that the parameter estimation can be updated/adjusted efficiently using real-time satellite observations in the GLA, leading to more accurate future predictions.

## 2 METHODOLOGY

**Physics-based simulator and satellite data** As stated in Papadopoulos & Pavlidou (2011), many geophysical variables, for instance, landscape slope, vegetation distribution and wind speed can sharply impact wildfire propagation. Taking these factors into consideration, various algorithms and softwares (Finney (1998); Alexandridis et al. (2008)) were developed to forecast wildfire spread on a regional scale. In this study, we made use of an operational CA model (Alexandridis et al. (2008)), to generate the training dataset for ROM and RNN surrogate modelling. In brief, this Cellular Automata (CA) approach uses square meshes to simulate the probabilistic spread of wildfires, relying on spatial data of local canopy density, canopy cover, landscape slope and wind speed [1]. A more detailed desription of the CA model can be found in the Appendix of this paper.

As for the observation data, MODIS provides a measurement of active fires on a daily basis with a resolution of 1km (Giglio et al. (2016)). VIIRS further improved fire detection capabilities with a finer spatial resolution. The observational records used in this paper are extracted from the polygon database to capture daily fire progression. The MODIS and VIIRS data are available about 2.5 hours after the satellite overpass, which allows a near real-time fire monitoring. To test the proposed approach in this work, we focus on three recent large wildfire events in California, namely the Pier fire, the Buck fire and the Chimney fire, from 2016 to 2020, as shown in table 1.

Table 1: Study areas of the three large wildfire events in California

| **Fire (Year)** | **latitude** | | **longitude** | | **area** |
|---|---|---|---|---|---|
| | North | South | West | East | |
| Chimney (2016) | 37.7366 | 37.5093 | -119.9441 | -119.7053 | $\approx 246\text{km}^2$ |
| Buck (2017) | 40.2558 | 40.1707 | -123.0791 | -122.9734 | $\approx 83\text{km}^2$ |
| Pier (2017) | 36.1909 | 36.0543 | -118.798698 | -118.616145 | $\approx 244\text{km}^2$ |

**Surrogate modelling** Based on snapshots of dynamical fire simulations, we first construct a low-dimensional latent space for two-dimensional (2D) state variables $\mathbf{u}_t$ which represent the current burned area via proper orthogonal decomposition (POD) or CAE, i.e.,

$$\tilde{\mathbf{u}}_t = \mathcal{E}(\mathbf{u}_t) \quad \text{and} \quad \mathbf{u}_t^{\text{rec}} = \mathcal{D}(\tilde{\mathbf{u}}_t), \tag{1}$$

where $\mathcal{E}, \mathcal{D}$ denote the encoder and the decoder, respectively and $\mathbf{u}_t^{\text{rec}}$ is the reconstruction at time $t$. A long short-term memory (LSTM) neural networks (NN) is then used to build sequence-to-sequence predictions following the simulation results projected/encoded in the reduced space, i.e.,

$$[\tilde{\mathbf{u}}_t, \tilde{\mathbf{u}}_{t+1}, ..., \tilde{\mathbf{u}}_{t+m_{\text{in}}-1}] \xrightarrow{\text{LSTM} \quad \text{prediction}} [\tilde{\mathbf{u}}_{t+m_{\text{in}}}, \tilde{\mathbf{u}}_{t+m_{\text{in}}+1}, ..., \tilde{\mathbf{u}}_{t+m_{\text{in}}+m_{\text{out}}-1}]. \tag{2}$$

Being a variant of RNN, LSTM manages to deal long term dependency, and vanishing gradient problems that standard RNN could not handle (Hochreiter & Schmidhuber (1997)). To integrate near real-time observations $\mathbf{y}_t$, the recently developed LA algorithm (Amendola et al. (2020)) is adapted in this study. As illustrated in figure 1, the preprocessed observation data are first encoded into the latent space using the encoder pre-trained on simulated data (i.e., $\tilde{\mathbf{y}}_t = \mathcal{E}(\mathbf{y}_t)$). Data assimilation then takes place in the reduced space to acquire an optimal approximation of the burned area relying on information embedded in $\tilde{\mathbf{u}}_t$ and $\tilde{\mathbf{y}}_t$. As shown in figure 1, this leads to an iterative process that improves the starting point of the next time-level forecast (Carrassi et al. (2018)), resulting in a more accurate prediction of the burned area.

[1] https://iftdss.firenet.gov/landing_page/

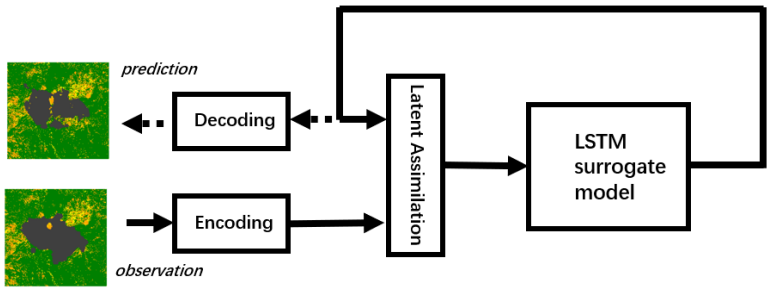



Figure 1: Flow chart of LSTM surrogate modelling with ROM and LA

**Parameter identification** Much effort has been devoted to parameter identification in fire spread models via inverse methods (Alessandri et al. (2021); Lautenberger (2013)). In the present paper, the CA model is mainly determined by four physical parameters $c_1, c_2, a, p_h$ as described in Alexandridis et al. (2008), where $c_1, c_2$ are related to the wind effect, $a$ is linked to the ground slope and $p_h$ is the standard spread probability (see a more detailed description in the Appendix). Data assimilation is a reference method for inverse problems, which aims to achieve an optimal compromise between model forecasts (or initial guess) and observational data. The standard formulation of variational DA in the assimilation window $\mathbf{T}_{\text{obs}}$ reads

$$\mathcal{J}(\mathbf{x}) = \frac{1}{2}||\mathbf{x} - \mathbf{x}_b||^2_{\mathbf{B}^{-1}} + \frac{1}{2}\sum_{t\in\mathbf{T}_{\text{obs}}} ||\mathbf{y}_t - \mathcal{H}_t(\mathbf{x})||^2_{\tilde{\mathbf{R}}_t{}^{-1}}, \quad \mathbf{x}_a = \arg\min_{\mathbf{x}}\Big(\mathcal{J}(\mathbf{x})\Big), \tag{3}$$

where $\mathbf{x}$ is the state vector ($\mathbf{x} = [c_1, c_2, a, p_h]$) for parameter estimation and $\mathbf{x}_b, \mathbf{x}_a$ are the prior and posterior estimations, respectively. $\mathcal{H}_t$ is the transformation function which maps the state variables to real-time observations. After acquiring the surrogate model, $\mathcal{H}_t$ is a composed of deep learning (DL) and ROM functions. Efficient machine learning (ML) forward prediction functions can be helpful for parameter identification. However, the minimisation of equation 3 requires the computation of the Hessian matrix (i.e.,Hess($\mathcal{J}$)) (Haben et al. (2011)) which can be cumbersome for ML functions due to its high complexity and the large number of parameters involved. To overcome this limit of current LA methods, we develop a new assimilation approach called GLA. Instead of assimilating the ML-based transformation function $\mathcal{H}$, local polynomial regressions (PRs) are performed in a neighbourhood of $\mathbf{x}_b$ to build a smooth local surrogate function to facilitate the computation of the Hessian matrix in the optimization loops. More precisely, Latin Hypercube Sampling (LHS)( Tang (1993)) is performed to build a PR learning ensemble $\{\mathbf{x}_b^q\}_{q=1..n_s}$ around the background state within certain range $r_s$. We then fit the ML model output by a local polynomial function, as shown in algorithm 1. Once the analysis state $\mathbf{x}_a$ is obtained, the forecasting of future burned area at time $t \in \mathbf{T}_{\text{pred}}$ can be performed,

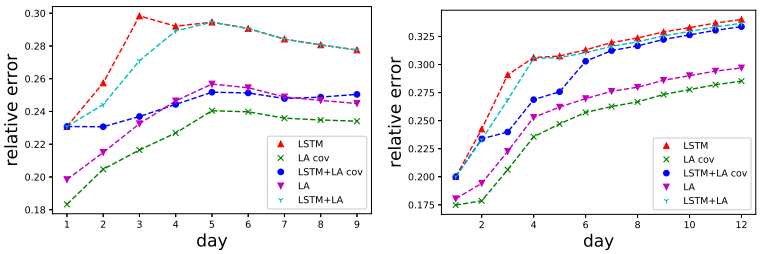
**Algorithm 1:** Generalised LA with local polynomial surrogate function

Inputs: $\mathbf{x}_b, \mathbf{B}, \{\tilde{\mathbf{R}}_t\}, \{\tilde{\mathbf{y}}_t\}, \{f_t^{\text{ML}}\}$ for $t \in \mathbf{T}_{\text{obs}}$
Paramters: $d_p$ (polynomial degree), $r_s$ (LHS range), $n_s$ (number of samplings), $P(d_p)$ denotes the space of polynomial functions of degree $d_p$

**for** $t \in \boldsymbol{T}_{obs}$ **do**
  $\{\mathbf{x}_b^q\}_{q=1..n_s} = \text{LHS Sampling}_{\{d^p, r_s, n_s\}}(\mathbf{x}_b)$
  $\tilde{\mathcal{H}}_t^p = \arg\min_{p\in P(d_p)}\big(\sum_{q=1}^{n_s} ||p(\mathbf{x}_b^q) - f_t^{\text{ML}}(\mathbf{x}_b^q)||_2\big)^{1/2}$
**end**

$\mathbf{x}_a = \arg\min_{\mathbf{x}}\Big(\frac{1}{2}||\mathbf{x} - \mathbf{x}_b||^2_{\mathbf{B}^{-1}} + \frac{1}{2}\sum_{t\in\mathbf{T}_{\text{obs}}} ||\tilde{\mathbf{y}}_t - \tilde{\mathcal{H}}_t^p(\mathbf{x})||^2_{\tilde{\mathbf{R}}_t{}^{-1}}\Big)$

outputs: $\mathbf{x}_a, \mathbf{u}_t^{\text{pred}} = \mathcal{D}\big(f_t^{\text{ML}}(\mathbf{x}_a)\big), \forall t \in \mathbf{T}_{\text{pred}}$

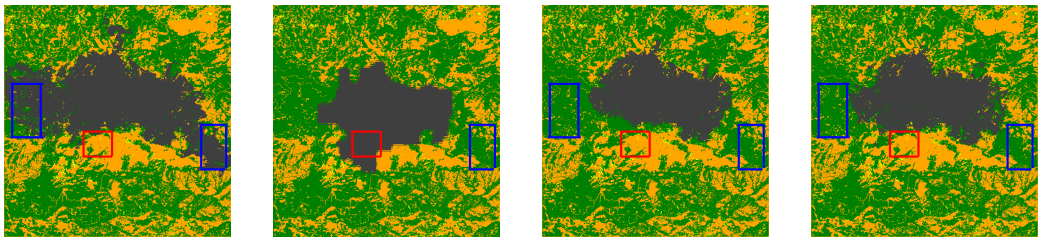



(a) Buck (b) Pier

Figure 2: Relative error of pure LSTM-based and assimilated prediction regarding satellite observations.

## 3 NUMERICAL RESULTS

**Surrogate modelling** We first tested the performance of the LSTM-based prediction model with ROM and LA applied to the Buck and the Pier fire. The evolution of the averaged relative root mean square error (R-RMSE) against time is shown in figure 2 where 'LSTM+LA' denotes the error of LSTM prediction from the previous assimilated results. Each step in CA and LSTM predictions is roughly equivalent to 30 minutes in real time while the satellite observations are of daily basis. The numerical results show that, with the help of error covariance tuning (Desroziers & Ivanov (2001)), the LA contributes to a more accurate reconstruction and future prediction in the original physical space for both fire events. As an example, the predicted and assimilated burned area for the Buck fire at day 3 (after ignition) is displayed in figure 3. The assimilated prediction (i.e.,'LSTM + LA cov') managed to reduce the forecasting error compared to pure LSTM results. The averaged online computational time for one CA/LSTM/LA step, carried out on a laptop CPU, is reported in table 2. For comparison purposes, we also display the approximately estimated computational time of widely used Flammap software (Finney (1998)). Thanks to reduced order surrogate modelling, the LSTM prediction is at least 1000 times faster in average compared to CA or Flammap simulations and the LA exhibits high efficiency, allowing near real-time wildfire prediction and updating.



(a) LSTM (b) obs (c) LSTM+LA (d) LA

Figure 3: Predicted and observed burned are for the Buck fire at day 3

Table 2: Averaged computational time for one time-step using different approaches

| Fire | CA | LSTM | LA | DI01 | Flammap |
|---|---|---|---|---|---|
| Buck | $2.96s$ | $5.26e^{-3}s$ | $4.26e^{-3}s$ | $2.12e^{-2}s$ | $5 \sim 20$s |
| Pier | $8.28s$ | $6.28e^{-3}s$ | $5.36e^{-3}s$ | $2.68e^{-2}s$ | $10 \sim 30$s |

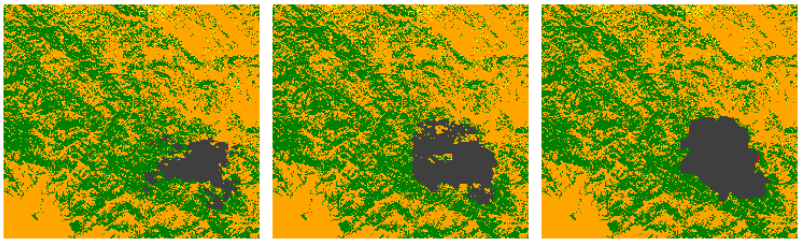
**Parameter estimation** In the study of Alexandridis et al. (2008), model parameters are fixed as

$$p_h = 0.58, \quad a = 0.078, \quad c_1 = 0.045, \quad c_2 = 0.131, \tag{4}$$

Table 3: Relative prediction error for the Chimney fire with assimilated parameters

| **Fire** | **latent space error** | | **full space error** | |
|---|---|---|---|---|
| | prior | posterior | prior | posterior |
| Chimney | 35.2% | 8.2% | 41.7% | 31.9% |

by fitting the observations in the specific fire event of application. In this work, these values are considered as the initial guess (i.e., $\mathbf{x}_b$ in equation 3) for parameter identification. Using the LHS technique, we included CA simulations with disturbed parameters in the computation of the ROM and the ML forward model. We show in table 3 the relative prediction error prior ($f_t^{\mathrm{ML}}(\mathbf{x}_b)$) and posterior ($f_t^{\mathrm{ML}}(\mathbf{x}_a)$) to the parameter identification in both the latent (encoded using a CAE) and the full physical space, evaluated on an unseen test data composed by 100 CA simulations. The prediction error has been considerably reduced in both the reduced-order and the full rank spaces thanks to GLA. The improvement in the full space is less significant due to the limited adaptive capacity of the CAE. We also illustrate in figure 4 the predicted burned area with and without parameter estimation, in comparison with satellite observation. GLA also manages to enhance the parameter estimation using satellite data.



(a) $f_t^{\mathrm{ML}}(\mathbf{x}_b)$ (b) $f_t^{\mathrm{ML}}(\mathbf{x}_a)$ (c) obs

Figure 4: Observed and predicted (with and without parameter assimilation) burned are for the Chimney fire at day 4

## 4 Discussion and future work

In this work, we have developed a fast ML- and ROM-based surrogate model for wildfire forecasting. As a first attempt at an efficient reduced order wildfire forecasting, the results presented in this paper highlight the potential of data-driven models for guiding fire suppression and evacuation strategies. We have also introduced a novel assimilation-based parameter estimation approach, named GLA, for ML fire spread models. Both the forward and inverses models have been tested in recent massive fire events in California with simulated and observed data, achieving a promising performance. More importantly, both methods (i.e., prediction and parameter estimation) can be adjusted efficiently by available observation data in near real-time. In terms of limitation, it is noticed that it is easier to correct overestimation (the zone framed by the blue square in figure 3) than underestimation (the zone framed by the red square) in LA for the surrogate model. In fact, since the CA depends heavily on the vegetation, the red zone has a very low probability to be burned in the training dataset. Therefore, future work should focus on improving the adaptive capability of current approaches by, for instance, enhancing the regularity of the AEs via Variational autoencoder (VAE) or domain adaption techniques (Peng et al. (2019)). Further study is also required to improve the data-driven model by learning from more complex fire simulations, which take into account the spotting spread and the probability of surface-to-crown transition in the proposed approaches.

## REFERENCES


Angelo Alessandri, Patrizia Bagnerini, Mauro Gaggero, and Luca Mantelli. Parameter estimation of fire propagation models using level set methods. *Applied Mathematical Modelling*, 92:731–747, 2021.

A. Alexandridis, D. Vakalis, C.I. Siettos, and G.V. Bafas. A cellular automata model for forest fire spread prediction: The case of the wildfire that swept through spetses island in 1990. *Applied Mathematics and Computation*, 204(1):191–201, 2008. ISSN 0096-3003.

Martin Ambroz, Karol Mikula, Marek Fraštia, and Marián Marčiš. Parameter estimation for the forest fire propagation model. *Tatra Mountains Mathematical Publications*, 75(1):1–22, 2018.

Maddalena Amendola, Rossella Arcucci, Laetitia Mottet, Cesar Quilodran Casas, Shiwei Fan, Christopher Pain, Paul Linden, and Yi-Ke Guo. Data assimilation in the latent space of a neural network, 2020.

C3S. Copernicus Climate Change Service (C3S) (2017): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate. Technical report, Copernicus Climate Change Service Climate Data Store (CDS), 2017. URL `https://cds.climate.copernicus.eu/cdsapp#!/home`.

Alberto Carrassi, Marc Bocquet, Laurent Bertino, and Geir Evensen. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, 9(5):e535, 2018.

César Quilodrán Casas, Rossella Arcucci, Pin Wu, Christopher Pain, and Yi-Ke Guo. A reduced order deep data assimilation model. *Physica D: Nonlinear Phenomena*, 412:132615, 2020.

Gongbo Chen, Yuming Guo, Xu Yue, Shilu Tong, Antonio Gasparrini, Michelle L Bell, Ben Armstrong, Joel Schwartz, Jouni JK Jaakkola, Antonella Zanobetti, et al. Mortality risk attributable to wildfire-related pm2· 5 pollution: a global time series study in 749 locations. *The Lancet Planetary Health*, 5(9):e579–e587, 2021.

Gerald Desroziers and Serguei Ivanov. Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. *Quarterly Journal of the Royal Meteorological Society*, 127 (574):1433–1452, 2001.

Matt Emschwiller, David Gamarnik, Eren C Kızıldağ, and Ilias Zadik. Neural networks and polynomial regression. demystifying the overparametrization phenomena. *arXiv preprint arXiv:2003.10523*, 2020.

Mark A Finney. *FARSITE, Fire Area Simulator–model development and evaluation*. Number 4. US Department of Agriculture, Forest Service, Rocky Mountain Research Station, 1998.

Louis Giglio, Wilfrid Schroeder, and Christopher O Justice. The collection 6 modis active fire detection algorithm and fire products. *Remote Sensing of Environment*, 178:31–41, 2016.

Stephen A Haben, Amos S Lawless, and Nancy K Nichols. Conditioning and preconditioning of the variational data assimilation problem. *Computers & Fluids*, 46(1):252–256, 2011.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

J Holladay. A note on the stone-weierstrass theorem for quaternions. In *Proc. Amer. Math. Soc*, volume 8, pp. 656–657, 1957.

Chris Lautenberger. Wildland fire modeling with an eulerian level set method and automated calibration. *Fire Safety Journal*, 62:289–298, 2013.

George D Papadopoulos and Fotini-Niovi Pavlidou. A comparative review on wildfire simulators. *IEEE systems Journal*, 5(2):233–243, 2011.

Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. *arXiv preprint arXiv:1911.02054*, 2019.

Omer San, Romit Maulik, and Mansoor Ahmed. An artificial neural network framework for reduced order modeling of transient flows. *Communications in Nonlinear Science and Numerical Simulation*, 77:271–287, 2019.

Olivier Talagrand. A posteriori evaluation and verification of analysis and assimilation algorithms. In *Workshop on Diagnosis of Data Assimilation Systems, 2 - 4 November 1998*, pp. 17–28, Shinfield Park, Reading, 1999. ECMWF, ECMWF.

Boxin Tang. Orthogonal array-based latin hypercubes. *Journal of the American statistical association*, 88(424):1392–1397, 1993.

Meet P Vadera, Adam D Cobb, Brian Jalaian, and Benjamin M Marlin. Ursabench: Comprehensive benchmarking of approximate bayesian inference methods for deep neural networks. *arXiv preprint arXiv:2007.04466*, 2020.

David R Weise and Gregory S Biging. A qualitative comparison of fire spread models incorporating wind and slope effects. *Forest Science*, 43(2):170–180, 1997.

# A APPENDIX

## CA FIRE SPREAD SIMULATOR

For the CA model in the present paper, the fire propagation towards 8 neighbour cells (as illustrated in figure 5) at a discrete time follows the probability,

$$P_{\text{bun}} = p_h(1 + p_{\text{veg}})(1 + p_{\text{den}})p_s p_w \tag{5}$$

where $p_{\text{veg}}$, $p_{\text{den}}$ , $p_s$ and $p_w$ are related to the local canopy density, canopy cover, landscape slope and wind speed/direction of the receiving cell, respectively. These physical fields for the corresponding study areas are obtained from the Interagency Fuel Treatment Decision Support System (IFTDSS) [2]. Based on the research of Weise & Biging (1997), $p_s$ and $p_w$ are modelled as:

$$p_s = \exp(a\theta_g), p_w = \exp(c_1 V_w) f_t, \quad \text{with} \quad f_t = \exp(V_w c_2(cos(\theta_w) - 1)) \tag{6}$$

where the ground slope angle $\theta_s$ is in function of elevation $E_i$ and cell length $l$,

$$\theta_g = \begin{cases} \tan^{-1}\left(\frac{E_1 - E_2}{l}\right) & \text{for adjacent cells} \\ \tan^{-1}\left(\frac{E_1 - E_2}{\sqrt{2}l}\right) & \text{for diagonal cells} \end{cases} \tag{7}$$

$V_w$ in equation 6 denotes the wind speed in $m/s$ and $\theta_w$ represents the angle between the wind direction and the potential fire propagation as shown in figure 5. The wind data is extracted from the dataset of C3S (2017). Since the grid scale is much sparser (i.e., $27km \times 27km$), wind speed and direction are considered as spatially constant as shown in figure 5.

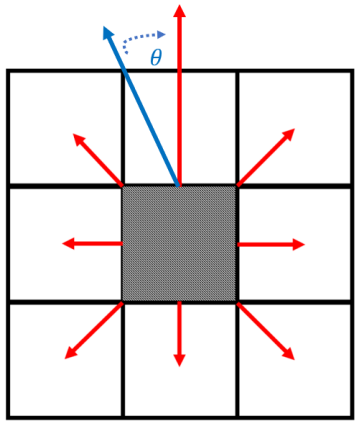



Figure 5: CA fire propagation modelling with wind effect

## THEORETICAL ANALYSIS OF GLA

For the sake of notation, we consider a general case of equation 3,

$$\mathcal{J}(\mathbf{x}) = \frac{1}{2}||\mathbf{x} - \mathbf{x}_b||^2_{\mathbf{B}^{-1}} + \frac{1}{2}||\mathbf{y} - \mathcal{H}(\mathbf{x})||^2_{\tilde{\mathbf{R}}^{-1}}, \quad \mathcal{J}^p(\mathbf{x}) = \frac{1}{2}||\mathbf{x} - \mathbf{x}_b||^2_{\mathbf{B}^{-1}} + \frac{1}{2}||\mathbf{y} - \mathcal{H}^p(\mathbf{x})||^2_{\tilde{\mathbf{R}}^{-1}}. \tag{8}$$

Since the LA is determined by the loss function $J(\mathbf{x})$, here we aim to provide an upper bound for the expected absolute and relative approximation error evaluated on the theoretical value of the reduced state $\mathbf{x}_{\text{true}}$ (also known as the true state in DA), i.e.,

$$\mathbb{E}\big(Jv(\mathbf{x}_{\text{true}}) - J(\mathbf{x}_{\text{true}})\big) \quad \text{and} \quad \frac{\mathbb{E}\big(J^p(\mathbf{x}_{\text{true}}) - J(\mathbf{x}_{\text{true}})\big)}{\mathbb{E}\big(J(\mathbf{x}_{\text{true}})\big)}. \tag{9}$$

[2] https://iftdss.firenet.gov/landing_page/

By definition,

$$J^p(\mathbf{x}) = \frac{1}{2}\Big(||\mathbf{x}-\mathbf{x}_b||^2_{\tilde{\mathbf{B}}^{-1}} + ||\tilde{\mathbf{y}}-\tilde{\mathcal{H}}(\mathbf{x})+\tilde{\mathcal{H}}(\mathbf{x})-\tilde{\mathcal{H}}^p(\mathbf{x})||^2_{\tilde{\mathbf{R}}^{-1}}\Big) \tag{10}$$

$$\leq \frac{1}{2}\Big(||\mathbf{x}-\mathbf{x}_b||^2_{\tilde{\mathbf{B}}^{-1}} + ||\tilde{\mathbf{y}}-\tilde{\mathcal{H}}(\mathbf{x})||^2_{\tilde{\mathbf{R}}^{-1}} + ||\tilde{\mathcal{H}}(\mathbf{x})-\tilde{\mathcal{H}}^p(\mathbf{x})||^2_{\tilde{\mathbf{R}}^{-1}}$$
$$+ 2||\tilde{\mathbf{y}}-\tilde{\mathcal{H}}(\mathbf{x})||_{\tilde{\mathbf{R}}^{-1}} \cdot ||\tilde{\mathcal{H}}(\mathbf{x})-\tilde{\mathcal{H}}^p(\mathbf{x})||_{\tilde{\mathbf{R}}^{-1}}\Big) \tag{11}$$

$$\leq \frac{1}{2}\Big(2J(\mathbf{x}) + ||\tilde{\mathcal{H}}(\mathbf{x})-\tilde{\mathcal{H}}^p(\mathbf{x})||^2_{\tilde{\mathbf{R}}^{-1}}\Big) + ||\tilde{\mathbf{y}}-\tilde{\mathcal{H}}(\mathbf{x})||_{\tilde{\mathbf{R}}^{-1}} \cdot ||\tilde{\mathcal{H}}(\mathbf{x})-\tilde{\mathcal{H}}^p(\mathbf{x})||_{\tilde{\mathbf{R}}^{-1}}. \tag{12}$$

Therefore,

$$\mathbb{E}\big(J^p(\mathbf{x}_{\text{true}})\big) \leq \mathbb{E}\big(J(\mathbf{x}_{\text{true}})\big) + \frac{1}{2}\mathbb{E}\big(||\tilde{\mathcal{H}}(\mathbf{x}_{\text{true}})-\tilde{\mathcal{H}}^p(\mathbf{x}_{\text{true}})||^2_{\tilde{\mathbf{R}}^{-1}}\big)$$
$$+ \mathbb{E}\big(||\tilde{\mathbf{y}}-\tilde{\mathcal{H}}(\mathbf{x}_{\text{true}})||_{\tilde{\mathbf{R}}^{-1}} \cdot ||\tilde{\mathcal{H}}(\mathbf{x}_{\text{true}})-\tilde{\mathcal{H}}^p(\mathbf{x}_{\text{true}})||_{\tilde{\mathbf{R}}^{-1}}\big). \tag{13}$$

It is often supposed in DA that both background and observation prior errors follow a centred Gaussian distribution, i.e.,

$$\mathbf{x}_b - \mathbf{x}_{\text{true}} \sim \mathcal{N}(0, \tilde{\mathbf{B}}), \quad \tilde{\mathbf{y}} - \tilde{\mathcal{H}}(\mathbf{x}_{\text{true}}) \sim (0, \tilde{\mathbf{R}}). \tag{14}$$

Therefore,

$$\sqrt{\tilde{\mathbf{B}}^{-1}}(\mathbf{x}_b - \mathbf{x}_{\text{true}}) \sim \mathcal{N}(0, \mathbf{I}_{\dim(\mathbf{x})}), \quad \sqrt{\tilde{\mathbf{R}}^{-1}}(\tilde{\mathbf{y}} - \tilde{\mathcal{H}}(\mathbf{x}_{\text{true}})) \sim \mathcal{N}(0, \mathbf{I}_{\dim(\tilde{\mathbf{y}})}). \tag{15}$$

As an important note, $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{R}}$ are real symmetric positive definite thus $\sqrt{\tilde{\mathbf{B}}^{-1}}$ and $\sqrt{\tilde{\mathbf{R}}^{-1}}$ are well-defined. This leads to

$$\mathbb{E}(||\mathbf{x}_{\text{true}} - \mathbf{x}_b||^2_{\tilde{\mathbf{B}}^{-1}}) = \mathbb{E}\left((\mathbf{x}_{\text{true}} - \mathbf{x}_b)^T \tilde{\mathbf{B}}^{-1}(\mathbf{x}_{\text{true}} - \mathbf{x}_b)\right) \tag{16}$$

$$= \mathbb{E}\left(\left(\sqrt{\tilde{\mathbf{B}}^{-1}}(\mathbf{x}_b - \mathbf{x}_{\text{true}})\right)^T \cdot \left(\sqrt{\tilde{\mathbf{B}}^{-1}}(\mathbf{x}_b - \mathbf{x}_{\text{true}})\right)\right) \tag{17}$$

$$= \mathbb{E}\left(||\sqrt{\tilde{\mathbf{B}}^{-1}}(\mathbf{x}_b - \mathbf{x}_{\text{true}})||_2^2\right) = \dim(\mathbf{x}) \tag{18}$$

As the same, $\mathbb{E}(||\tilde{\mathbf{y}} - \tilde{\mathcal{H}}(\mathbf{x})||^2_{\tilde{\mathbf{R}}^{-1}}) = \dim(\tilde{\mathbf{y}})$, resulting in

$$\mathbb{E}\big(J(\mathbf{x}_{\text{true}})\big) = \dim(\mathbf{x}) + \dim(\tilde{\mathbf{y}}). \tag{19}$$

A similar conclusion can be found in the work of Talagrand (1999), proved via establishing the Mahalanobis norm. Now we concentrate on the other terms of equation 13. As a matter of fact, the observation error $||\tilde{\mathbf{y}} - \tilde{\mathcal{H}}(\mathbf{x}_{\text{true}})||_{\tilde{\mathbf{R}}^{-1}}$ is only related to instrument, representation or encoding errors. While the approximation error $||\tilde{\mathcal{H}}(\mathbf{x}) - \tilde{\mathcal{H}}^p(\mathbf{x})||_{\tilde{\mathbf{R}}^{-1}}$ is only caused by the polynomial regression where the real observation $\mathbf{y}$ is not involved. Therefore, we suppose that $||\tilde{\mathbf{y}} - \tilde{\mathcal{H}}(\mathbf{x}_{\text{true}})||_{\tilde{\mathbf{R}}^{-1}}$ is uncorrelated to $||\tilde{\mathcal{H}}(\mathbf{x}) - \tilde{\mathcal{H}}^p(\mathbf{x})||_{\tilde{\mathbf{R}}^{-1}}$, leading to

$$\mathbb{E}\big(||\tilde{\mathbf{y}} - \tilde{\mathcal{H}}(\mathbf{x}_{\text{true}})||_{\tilde{\mathbf{R}}^{-1}} \cdot ||\tilde{\mathcal{H}}(\mathbf{x}_{\text{true}}) - \tilde{\mathcal{H}}^p(\mathbf{x}_{\text{true}})||_{\tilde{\mathbf{R}}^{-1}}\big)$$
$$= \mathbb{E}\big(||\tilde{\mathbf{y}} - \tilde{\mathcal{H}}(\mathbf{x}_{\text{true}})||_{\tilde{\mathbf{R}}^{-1}}\big) \cdot \mathbb{E}\big(||\tilde{\mathcal{H}}(\mathbf{x}_{\text{true}}) - \tilde{\mathcal{H}}^p(\mathbf{x}_{\text{true}})||_{\tilde{\mathbf{R}}^{-1}}\big) = 0. \tag{20}$$

Now the only remaining task is to bound the polynomial regression error. For this, we rely on the recent theoretical analysis of Emschwiller et al. (2020), which proves that for approximating a teacher NNs via polynomial regression,

$$N' = \begin{cases} = d^{O(L/\epsilon')^L} & \text{for the ReLU activation function} \\ \\ = d^{O(log^L(L/\epsilon'))} & \text{for the Sigmoid activation function,} \end{cases} \tag{21}$$

where $N'$ is the required number of samples in the training dataset, $d$ is the dimension of the input vector, $L$ is the number of NNs layers and $\epsilon'$ is the target relative prediction error (in our case $\epsilon = \Big(||\tilde{\mathcal{H}}(\mathbf{x}) - \tilde{\mathcal{H}}^p(\mathbf{x})||_2 / ||\tilde{\mathcal{H}}(\mathbf{x})||_2\Big) \leq \epsilon'$). Since we are looking for a boundary of the regression error $\epsilon$,

$$N' = d^{\Big(c(L/\epsilon')^L\Big)} \quad \text{where} \quad c \text{ is a real constant} \Leftrightarrow \quad \log_d N' = c(L/\epsilon')^L \tag{22}$$

$$\Leftrightarrow \quad \Big(\frac{\log_d N'}{c}\Big)^{1/L} = L/\epsilon' \Leftrightarrow \quad \epsilon \leq \epsilon' = L\Big(\frac{c}{\log_d N'}\Big)^{1/L} \tag{23}$$

$$\Leftrightarrow \quad ||\tilde{\mathcal{H}}(\mathbf{x}) - \tilde{\mathcal{H}}^p(\mathbf{x})||_2 \leq L\Big(\frac{c}{\log_d N'}\Big)^{1/L} ||\tilde{\mathcal{H}}(\mathbf{x})||_2. \tag{24}$$

Since the $L^2$ norm and the $||.||_{\tilde{\mathbf{R}}^{-1}}$ norm are equivalent (in a finite dimensional vector space), one can extend the boundary in equation 24 to

$$\begin{aligned}\mathbb{E}\big(J^p(\mathbf{x}_{\text{true}})\big) &\leq \mathbb{E}\big(J(\mathbf{x}_{\text{true}})\big) + \frac{1}{2}\text{cond}(\tilde{\mathbf{R}})L^2\Big(\frac{c}{\log_d N'}\Big)^{2/L}\mathbb{E}\big(||\tilde{\mathcal{H}}(\mathbf{x}_{\text{true}})||^2_{\tilde{\mathbf{R}}^{-1}}\big) \\ &\leq \dim(\mathbf{x}) + \dim(\tilde{\mathbf{y}}) + \frac{1}{2}\text{cond}(\tilde{\mathbf{R}})L^2\Big(\frac{c}{\log_d N'}\Big)^{2/L}\mathbb{E}\big(||\tilde{\mathcal{H}}(\mathbf{x}_{\text{true}})||^2_{\tilde{\mathbf{R}}^{-1}}\big).\end{aligned} \tag{25}$$

where $\text{cond}(\tilde{\mathbf{R}})$ is the condition number of the $\tilde{\mathbf{R}}$ matrix. We have thus obtained an upper bound of $\mathbb{E}\big(J^p(\mathbf{x}_{\text{true}})\big)$ and $\mathbb{E}\big(J^p(\mathbf{x}_{\text{true}})\big) - \mathbb{E}\big(J(\mathbf{x}_{\text{true}})\big)$ unrelated to the local polynomial surrogate function $\tilde{\mathcal{H}}^p$. Furthermore, in the case where the target NNs is fixed and we have infinite local training data for the polynomial surrogate model,

$$\mathbb{E}\big(J^p(\mathbf{x}_{\text{true}}) - J(\mathbf{x}_{\text{true}})\big) \xrightarrow{N' \to +\infty} 0. \tag{26}$$

This is consistent with the Stone–Weierstrass theorem which proves the fact that all continuous functions defined on a closed interval can be approximated as closely as desired by a polynomial function (Holladay (1957)).