Machine Learning with Applications 5 (2021) 100053

Contents lists available at ScienceDirect

Machine Learning with Applications

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

Convection indicator for pre-tactical air traffic flow management using neural networks

Aniel Jardines [a,*], Manuel Soler [a], Alejandro Cervantes [b], Javier García-Heras [a], Juan Simarro [c]

[a] Department of Bioengineering and Aerospace Engineering, Universidad Carlos III de Madrid, Avda. Universidad 30, 28911 Leganes, Spain
[b] Department of Computer Science and Engineering, Universidad Carlos III de Madrid, Avda. Universidad 30, 28911 Leganes, Spain
[c] Agencia Estatal de Meteorología (AEMET), Valencia, Spain

ARTICLE INFO

Keywords:
Thunderstorms
Air traffic management
Numerical weather prediction
Satellite images
Machine learning

ABSTRACT

Convective weather is a large source of disruption for air traffic management operations. Being able to predict thunderstorms the day before operations can help traffic managers plan around weather and improve air traffic flow management operations. In this paper, machine learning is applied on data from satellite storm observations and ensemble numerical weather prediction products to detect convective weather 36 h in advance. The learning task is formulated as a binary classification problem and a neural network is trained to predict the occurrence of storms. The neural network results are used to develop a probabilistic based convection indicator capable of outperforming existing convection indicators found in the literature. Lastly, applications of the neural network based indicator in an air traffic management setting are presented.

1. Introduction

Convective weather is a well known aviation hazard; turbulence, wind shear, lighting, and hail are elements arising in thunderstorms that can be catastrophic for aircraft. In Europe, convective weather, i.e. thunderstorms typically occur in the summer and coincide with a period of high air traffic demand on the airspace system. This combination of bad weather and high demand causes significant disruption to air traffic management operations resulting in delays throughout the network. In 2018, 25% of the total delay in the European airspace was attributed to adverse weather, resulting in a total of 4.8 million minutes, the majority can be attributed to convective weather (EUROCONTROL, 2019). Using the estimated rate of 100€ per minute of delay (Cook & Tanner, 2015), the costs associated with the weather delay in 2018 can be quantified at 0.48 billion euros.

A key reason why thunderstorm phenomena are so disruptive is the difficulty of forecasting their birth and evolution. While some meteorological conditions are required for thunderstorm formation and can be forecast in advance, the specific location and timing of convective initiation triggers is harder to identify. As a consequence, thunderstorm prediction is usually performed using nowcasting. Nowcasting are short term predictions, typically 1–3 h, based on extrapolation of sensor data such as Doppler radars or satellite (Wilson et al., 1998). However, extrapolation degrades rapidly as the forecasting horizon increases. One nowcasting system of particular interest for aviation is the Corridor Integrated Weather System (CIWS) (Evans & Ducot, 2006), which is in use in the US.

Due to the poor prediction precision of convective weather at time horizons greater than 3 h, air navigation service providers and airlines typically do not make strategic modifications to their operational plans, instead preferring to make tactical adjustments in real-time according to the evolving weather situation. This reactive approach in handling convective weather events is not conducive to coordination among multiple Air Navigation Service Providers (ANSP) and leads to inefficiency in the system.

The process of Air Traffic Flow Management (ATFM) aims at minimizing network disruptions in the system by matching the airspace and airport capacity with the varying levels of traffic demand to ensure safety and efficiency throughout the airspace system. ATFM is a coordinated effort between multiple stakeholders including the Network Manager, national ANSPs, and aircraft operators. ATFM is a multi-phase iterative process beginning months before the day of operations.

The pre-tactical phase of ATFM focuses on measures to be applied at least one day prior to the day of operations. In this stage, analysis is performed to refine capacity and demand estimates, and assess ATFM measures. The outcome of this phase is a plan for the day of operations, known as the ATFM Daily Plan (ADP) in Europe.

While weather condition are considered during this phase of ATFM, the weather information available for input to the ATFM Daily Plan

[*] Corresponding author.
E-mail addresses: ajardine@ing.uc3m.es (A. Jardines), masolera@ing.uc3m.es (M. Soler), acervant@inf.uc3m.es (A. Cervantes), gcarrete@ing.uc3m.es (J. García-Heras), jsimarrog@aemet.es (J. Simarro).

https://doi.org/10.1016/j.mlwa.2021.100053
Received 2 November 2020; Received in revised form 25 May 2021; Accepted 29 May 2021
Available online 9 June 2021

is limited. In Europe, EUROCONTROL's Network Operations Portal provides a Daily Network Weather Assessment, a document containing a brief written description of the general weather outlook for the Network, and severe weather alerts for en route airspaces and airports. The weather assessment also contains a series of static maps providing forecasts of temperature, winds and precipitation for the day. While this daily product is useful in providing some awareness of the meteorological conditions for the day, it fails to capture evolving weather phenomena such as convection. In order to effectively minimize the disruptions on the network, traffic managers require high confident convective weather forecast with sufficient lead time.

In order to extend the lead time in thunderstorm prediction it is necessary shift away from nowcasting methods and exploit the advances in Numerical Weather Prediction (NWP) tools. NWPs use computer simulations to model the atmospheric processes at a computational grid. The fluid motion and thermodynamic characteristics of the atmosphere are modeled using partial differential equations, capturing interactions among neighboring grid cells and calculating a broad set of atmospheric parameters for each grid cell. These NWP products are able to predict the state of the atmosphere multiple days into the future with fairly good accuracy. Indeed, the majority of the weather forecast we use in our daily lives rely on NWPs. However, NWPs have not traditionally been used for thunderstorm prediction because the size and lifespans of thunderstorms are small compared with the spatiotemporal resolution of medium-range NWP models.

Advances in weather science and high performance computing have greatly improved the prediction skill of NWPs in recent years. In our research we set out to leverage these improvements and machine learning techniques to predict thunderstorms using NWPs at timescales (greater that 24 h) required for the pre-tactical phase air traffic flow management.

At shorter time horizons, machine learning and NWPs have been used successfully to improve nowcasting of thunderstorms. In Mecikalski et al. (2015), machine learning techniques were applied on satellite data to improved their nowcasting algorithm's ability to predict which cloud objects would display convective initiation within the hour. Also, in Li et al. (2019) machine learning techniques are applied to Doppler radar images to predict gale force winds. Also, in Khandan et al. (2018) a Random Forest is used to predict convection initiation for the next 6 h from satellite and NWP data. However, predictions at these time scales are not compatible with pre-tactical ATFM operations.

Machine learning has also been applied on NWP data to predict thunderstorms for longer time horizons. In Šaur (2017), NWP and historical weather data are used to train a back-propagation algorithm to predict convective precipitation that may cause to flash floods over the Zlin region of Czech Republic up to 24 h in advance. In Collins and Tissot (2015), a deep-learning neural network model is developed using cloud to ground lightning data to predict the occurrence of thunderstorms in certain regions of Texas, US within 2 h time steps at time horizons up to 15 h. Random Forest has also been applied on NWP to predict the probability of lightning strike over the Alaskan tundra (He & Loboda, 2020). In Simon et al. (2018), thunderstorm occurrence within a 6 h period is predicted over the European eastern Alps up to 5 days in advance using generalized additive models (GAMs) and gradient boosting with cloud-to-ground lightning data. Convolutional Neural Networks have also been applied on NWP products to predict multiple types of convective weather within a 6 h period up to 72 h in advance (Zhou et al., 2019). While these works have been successful in using machine learning to predict convective weather, their specific applications did not require spatial–temporal resolution nor the continental scale geographic domain necessary for pre-tactical ATFM application. While works predicting convective events with high spatial–temporal resolution do exist (Baldauf et al., 2011; Spiridonov et al., 2020), they rely on physics-based computational fluid dynamic models rather than machine learning, and are limited in their geographical domain.

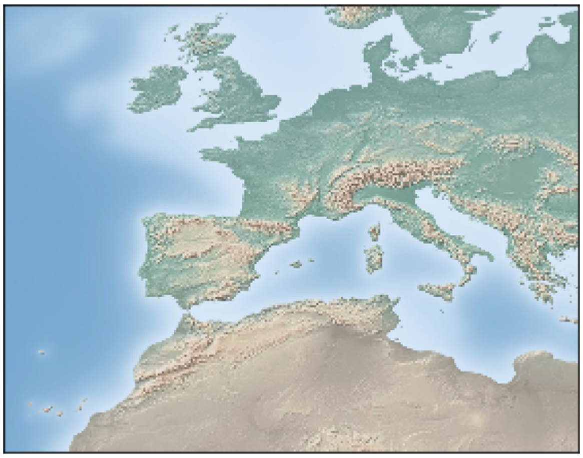


Fig. 1. Geographical domain of forecast and observational weather data.

In this paper we apply machine learning to predict thunderstorm occurrence over a large portion of western Europe in hourly time steps at time horizons up to 36 h. An ensemble NWP with 0.25 degree spatial resolution and satellite observations from the EUMETSAT NWC-SAF Rapid-Development Thunderstorm product are used to train a neural network to provide the likelihood of convective weather. To the authors' knowledge, the use of satellite storm data is novel approach, previous works using machine learning to predict convective weather has relied on cloud-to-ground lightning.

Model results are used to create a convection indicator that enables the consideration of thunderstorms during the pre-tactical phase of ATFM. The novel indicator is compared with an existing convection indicator found in the literature. The remainder of this paper is organized as follows. Section 2 presents an overview of the data used, while details of the neural network at provided in Section 3. Next, results are presented in Section 4, followed by examples of model application within an ATFM context in Section 5. Finally, a summary is provided in Section 6 where conclusions and future work are discussed.

2. Weather data

Convection is a vertical phenomena in the atmosphere created by the uneven heating of the Earth's surface due to solar radiation. Heat from the Earth's surface warms the air directly above it, causing the air to expand, becoming less dense than the surrounding air, and creating thermal columns of rising air. If moisture is also present, the warm moist air will rise and in the processes cool and condense. If sufficient instability is present in the atmosphere, this process can form extensive towering cumulonimbus clouds creating ideal conditions for thunderstorms. Convective storms can become quite extensive and be observed from space.

In developing the convection prediction model, data from ensemble NWP forecasts and satellite thunderstorms observations are used. Given the lead times required for pre-tactical ATFM, the model input is provided by ensemble NWP forecasts, as these are available 36 h in advance. Satellite image data is used for training and evaluation of the model as it provides an accurate representation of convective events. The data used is from June 2018 with a geographical domain covering vast portions of western Europe and northern Africa as seen in Fig. 1.

2.1. Ensemble NWP

Ensemble probabilistic forecasting is a technique used to provide an estimate of the uncertainty associated with predictions of the atmosphere. Rather than forecasting one future scenario as in traditional

NWPs, multiple future scenarios are created, using a variety of techniques including perturbing initial conditions, running multiple models, or using different combinations of physical parameterization schemes. The perturbation techniques are inline with the observational errors in the current state of the atmosphere. An assumption in using ensemble forecasts is that the probability of occurrence for each member is equally likely. A priori, there is no way of knowing which members will more closely resemble actual conditions. Furthermore, one ensemble member may be closest to the truth at a given geographical location, but this need not be the case at another location (Palmer et al., 2006). The spread of the members will reflect the predictability of the atmosphere, with a larger deviation between members indicative of a less predictable atmosphere. The goal of the ensemble system it to capture reality within the range of predictions. The ensemble NWP data used in this research comes from European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS). The EPS product is comprised of a control member, using the most accurate estimate of the initial conditions, plus 50 perturbed members. The forecasts are released twice a day at 00:00 and 12:00 UTC and provide a prediction of the weather up to 15 days ahead (Molteni et al., 1996).

In developing our model we use data from the 50 perturbed members, focusing on the forecast provisions for the next 36 h in 1 h steps. The spatial resolution of the EPS perturbed members is a quarter of a degree in latitude and longitude, this equates to roughly 15 nautical miles between grid points.

In selecting the NWP parameters for training or model we chose those that could best capture the physics of convective weather and thunderstorms. Our selection was guided based on the principle that thunderstorms are most likely to occur under the following conditions (Oxf, 2015):

- Lifting force or trigger mechanism to produce early saturation of air. In convective storms, this trigger action is typically caused by heat from the earth's surface causing moist air to rise.
- Sufficient moisture in the atmosphere to form and maintain the cloud.
- Atmospheric instability determined by the vertical temperature profile or lapse rate.

With these conditions in mind, 18 NWP parameters from the EPS were selected to train the NN model. Besides these 18 NWP parameters, we also included additional parameters to train our model. The parameter *hour of the day* was added to account for the weather patterns that occur throughout the diurnal cycle. The time horizon or *range of forecast* was also added, this parameter describes how far into the future a prediction is made. We hypothesized that our model may give more weight to certain parameters based on the range. Additionally, we also trained the model with the Convective available potential energy (cape) parameter values from the three previous time steps of the ensemble product because large values of cape correlate with time periods leading up to the storm, rather than during the storm itself. This provided us with a total of 23 input parameters (18 ECMWF parameters + 1 Hour of day + 1 Range of forecast + 3 time lagged CAPE) to train our model. The complete list of parameters and abbreviation is provided in Table 1.

2.2. Satellite data

Geostationary satellites with orbital periods that match the Earth's rotation allow for continuous observation of specific regions. Visual and infrared satellite imagery captures vital information regarding cloud cover, water vapor and temperature that allow for monitoring and tracking of weather.

The Rapid-Development Thunderstorm (RDT) product was developed by Météo-France within the EUMETSAT NWC-SAF framework. The RDT algorithm employs primarily geostationary satellite data to provide information about clouds related to significant convective systems from the mesoscale (200–2000 km) down to tenths of kilometers (Lee et al., 2020). The RDT product outputs storm information on a 15 min interval. For each cloud cell, the RDT product defines a series of parameters capturing the location, shape, cloud top, movement, severity, and life cycle phase. Despite the rich characterization of thunderstorms by the RDT product, only the location and shape information of convective cells is used to create the labeled "truth" training data required for supervised learning type problems.

Table 1
Total list of parameters used to train model.

Parameter	Short name
2 meter dewpoint	2d
2 meter temperature	2t
Convective available potential energy	cape
Convective available potential energy 1 h before	cape-1
Convective available potential energy 2 h before	cape-2
Convective available potential energy 3 h before	cape-3
Convective inhibition	cin
Convective precipitation	cp
Convective rain rate	crr
Height of convective cloud top	hcct
Hour of day	hour
K index	kx
Large scale precipitation	lsp
Large scale rain rate	lsrr
Surface latent heat flux	slhf
Surface pressure	sp
Surface sensible heat flux	sshf
Range of forecast	range
Total cloud cover	tcc
Total column water	tcw
Total column water vapor	tcwv
Total totals index	totalx
Geopotential	z

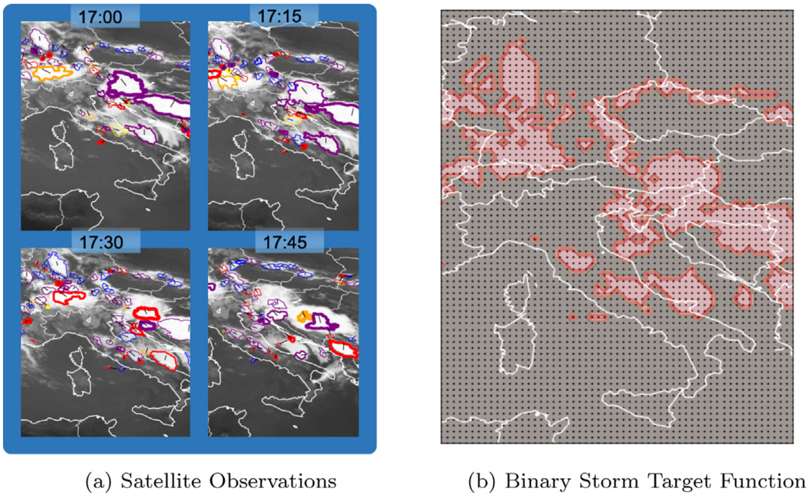
Table 2
Dates used for training, validation and testing.

Training		Validation	Test
Jun-01	Jun-02	Jun-03	Jun-04
Jun-05	Jun-06	Jun-07	Jun-08
Jun-09	Jun-10	Jun-11	Jun-12
Jun-13	Jun-14	Jun-15	Jun-16
Jun-17	Jun-18	Jun-19	Jun-20
Jun-21	Jun-22	Jun-23	Jun-24
Jun-25	Jun-26	Jun-27	Jun-28
Jun-29	Jun-30		

2.3. Data integration

Train, validation and test data sets are created by integrating the NWP forecast and the RDT satellite images. By projecting the NWP grid onto the higher resolution satellite images and identifying the grid points within the RDT storm polygons it is possible to express the data using a common spatial resolution of .25 degree x .25 degree. To reconcile the differences in the temporal resolution, 1 h for the NWP predictions versus 15 min for the RDT observations, a grid point is classified as convective if a storm observation is present during any of the four observations instances within the hour. In this way a binary training target function is constructed representative of storm cell occurrence at a grid location within the hour. Fig. 2 shows an example of how four satellite images are processed to establish from the target function.

Because we are interested in a time horizon of 36 h, and forecasts are released every 12 h, different range forecasts valid for the same time are used to train, validate and test the model. Having data at varying forecast ranges will allow us to analyze how the forecast degrades with increasing time horizon.



(a) Satellite Observations

(b) Binary Storm Target Function

Fig. 2. RDT satellite observations and resulting target function for thunderstorms occurring at 17:00 on June 8th, 2018.

3. Methodology

Before undertaking this study, a series of smaller preliminary case studies was conducted on a limited data set. Multiple machine learning models were developed using storm observations from the RDT product and one ensemble member from the ECMWF EPS product. The NWP product consisted of a 0.25 degree spatial grid, a 3 h time step and 48 h forecast range. Eight days of worth of data, June 4th–June 11th, were split to create train, validation and test data sets. Results from the study showed that a Neural Network architecture was superior than other methods including Logistic Regression, Decision Tree, and Random Forrest. It was assumed that the advantage in prediction skill for the NN method would hold true on a data set incorporating all 50 members and smaller time step of 1 h. A Convolutional Neural Networks (CNN) architecture was also considered, however given the high resolution of our data set; hourly time steps, 36 h forecast range, .25 degree spatial resolution, a geographical domain of western Europe, and the 50 ensemble members, using a CNN approach supposed a significant increase in computational cost. While a CNN methodology may be able to capture correlation among neighboring grid points, it is assumed that the physical process behind these interactions is already captured to some degree within the NWP model. As a result, for the purposes of training the neural network model, the data from each grid point is assumed to be an independent data sample. Also, during training each forecast member is treated individually. In this way, during training the model sees 50 separate data samples from each grid point in the ensemble forecast. The intention is for the model to benefit from ensemble distribution of parameters and adjust the neural network node weights accordingly.

The convection predictive model is trained to predict the probability of a thunderstorm occurring at given location and time. Only the 23 parameters defined in Table 1 are considered in making a prediction. In evaluating the model, each forecast member is evaluated separately, and results are averaged over the 50 members to obtain a probability.

For this study an integrated data set of EPS forecast and RDT observations covering the month of June 2018 is used. From the 30 days in June, 16 days are selected for training, 7 days for validation and 7 for testing, exact dates used for each data subset can be seen in Table 2. This partition was preferred over a sequential split to ensure sufficient convective samples in each subset. It is acknowledged that having a test data set embedded within the training data could introduce look-ahead bias in our results, however considering the NWP forecast is provided in hourly time steps and that the lifespan of convective events is on the order of hours, each day is treated independently. We assume that the convective events occurring on a specific day are independent from those occurring on a separate day. While temporal correlations exist in the atmosphere over consecutive days, we assume these correlations are more likely to be inherent within the NWP input than learned by the model.

3.1. Neural network model

The learning task of predicting convective weather was formulated as a binary classification problem. Based on the 23 inputs derived from the NWP forecast our model was trained to classify a grid-point as either convective (class 1) or not-convective(class 0). It is important to note that the model does not consider the latitude–longitude of the grid-point, providing a location-independent prediction based only on the physical NWP parameters.

A Multi-layer Perceptron (MLP) neural network was created using the python keras library to fit the data. The 23 NWP features were normalized using a standard scaler function before fitting the model to account for the order of magnitude differences between the values. The model consisted of the input layer with 23 nodes, two hidden layers of 16 nodes each and the output layer containing one node. The nodes in the hidden layers of the model used a Rectified Linear Unit activation function, while the node in the output layer used a Sigmoid activation function. By having a Sigmoid output, the model predicts a value between 0 and 1 instead of binary. This output value is representative of the confidence the sample is convective (class 1). Additionally, during training dropout layers of fraction 0.2 were introduced after each hidden layer. Dropout is a technique to prevent over-fitting of the model by randomly ignoring a fraction of the nodes during each iteration of training, in effect reducing the interdependent learning between neurons (Srivastava et al., 2014). In Fig. 3 a schematic representation shows the model architecture and data process from EPS data to convection indicator.

It is important to note that the data was highly imbalanced, with roughly 90% of the samples belonging to the non-convective class. To account for this imbalance, class weighting factors were applied during training. By implementing a class weighting factor, the binary cross-entropy loss function used for training assigned higher values to instances of the minority convective class. This reduces the impact of the majority class in the loss function, preventing the generation of models that basically predict the majority (non-convective) class for all samples. The weighted binary cross entropy loss function is defined in Eq. (1), where w_i is the weight factor for each class, t is the truth value of 0 or 1, and p is the probability of the sample belonging to the convective class.

$$CE = -w_i \sum_{i=0}^{C'=1} t_i log(p_i) = -w_i[tlog(p) + (1-t)log(1-p)] \tag{1}$$

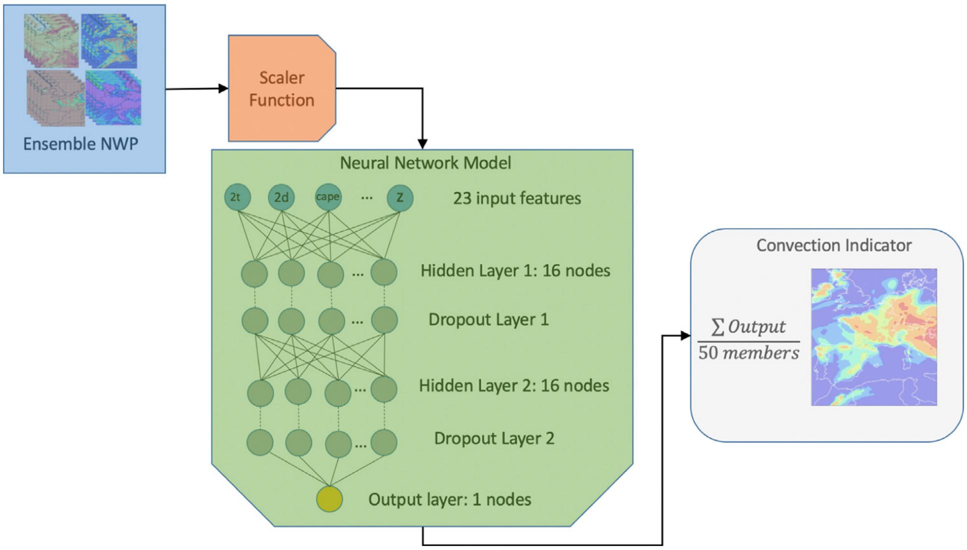


Fig. 3. Schematic of neural network model, showing data flow from ensemble NWPs to convection indicator.

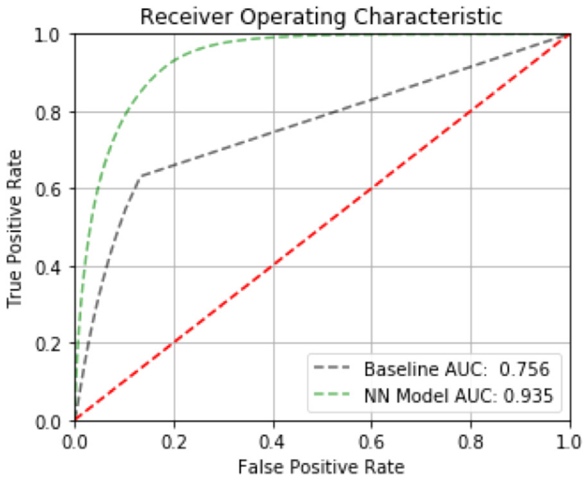


Fig. 4. ROC curve comparing performance of baseline and neural network indicators for entire test data set.

4. Results

In this section we present the results of the neural network model for the seven days in our test data set. For comparison we also present the results from an existing NWP based convection indicator, in our discussion we will refer to this indicator as the baseline. A brief description of the baseline indicator is provided below.

4.1. Baseline indicator description

An existing convection indicator used within an aviation context was found in the literature (González-Arribas et al., 2019). The indicator relies on two parameters from a numerical weather prediction product; Total Totals Index and Convective Precipitation . Total Totals Index ($totalx$) is the temperature and moisture gradient in the lower levels atmosphere and an indication of instability. Convective Precipitation (cp) is the accumulated water that falls to the Earth's surface that is generated by convection. Convection can be defined as an area where there is atmospheric instability and precipitation. Thus we can evaluate each point of the numerical weather prediction model for convection using the logistic expression in Eq. (2).

$$Convection = (totalx > TT_{TH}) \wedge (cp > 0) \tag{2}$$

where TT_{TH} is defined as the Total Totals Index threshold value. This threshold value can be associated with various levels of convection. The correlation between threshold value and convection severity is provided below.

- 44–45 isolated moderate thunderstorms
- 46–47 scattered moderate / few heavy thunderstorms
- 48–49 scattered moderate / few heavy / isolated severe thunderstorms
- 50–51 scattered heavy / few severe thunderstorms and isolated tornadoes
- 52–55 scattered to numerous heavy / few to scattered severe thunderstorm/few tornadoes
- 55+ numerous heavy / scattered severe thunderstorms and scattered tornadoes

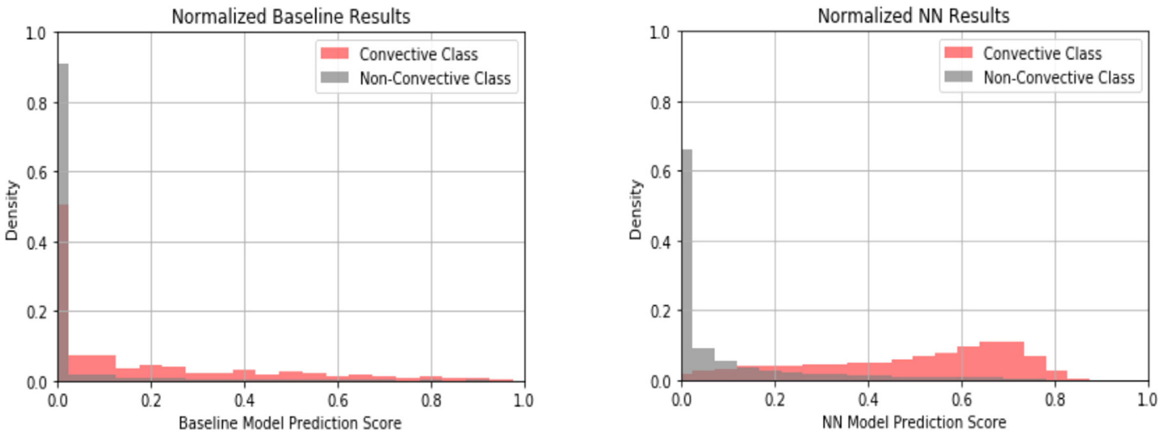
Eq. (2) is used to evaluate each grid point of the NWP. If both conditions in the logistic expression are met the grid point location is classified as convective (1), if the conditions are not met the location is classified as non-convective (0). These binary values are then averaged over the 50 ensemble members to provide the final Baseline Indicator score.

In our application of the baseline indicator we will assume a Totals Total Index threshold value of 44, and rather than using cp, which gives an accumulated value of convective precipitation since the forecast release, we will utilize the convective rain rate (crr). It is important to note that while cp is an accumulated parameter, crr is considered an instantaneous parameter, and not representative of the rain rate over the entire time step. Nonetheless, using the parameter crr instead of cp will better account for convective weather at discrete time steps in the forecast. The expression used to calculate the baseline convection indicator is provided in Eq. (3).

$$Convection = (totalx > 44) \wedge (crr > 0) \tag{3}$$

4.2. Model comparison

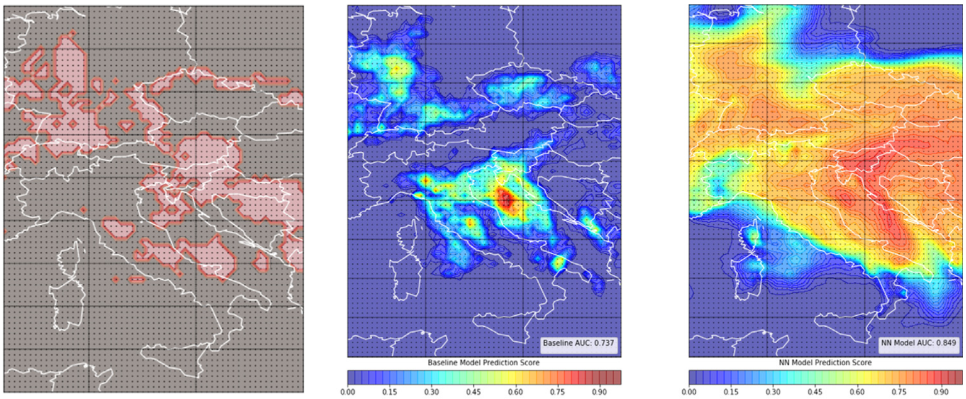
The effectiveness of our NN convection indicator is compared with the baseline indicator using a receiver operating characteristic (ROC) curve. A ROC curve is a technique used to evaluate binary classifiers by plotting the sensitivity, or true positive rate (TPR), against (1-specificity), or the false positive rate (FPR), for various threshold settings (Mandrekar, 2010). The TPR provides the probability of detection, and the FPR provides the probability of false alarm. The ideal



(a) Baseline Class Distribution

(b) NN Model Class Distribution

Fig. 5. Normalized histograms showing the class distributions by predictive score of baseline and neural network models for test data set.



(a) Target Function

(b) Baseline, AUC: 0.737

(c) NN Model, AUC: 0.849

Fig. 6. Binary thunderstorm target function compared with baseline and neural network model predictions for 17:00 UTC on June 8th, 2018 (Forecast range:17 h).

classifier would have a curve close to the upper left corner of the graph and maximizing the area under the curve (AUC). The diagonal line dividing the ROC space represents a random classifier, points above the line indicate the classifier performs better than random guessing. Model performance can be tuned by selecting a threshold value on the curve, which leads to a specific pair (FPR, TPR). In practice, the selection of a threshold value is linked with the amount of risk a user is willing to assume. A low threshold would increase the likelihood of capturing the thunderstorms, while also overestimating their presence in the airspace (false alarm). However, choosing a high threshold value would minimize the false alarm rate, at the risk of missing a portion of the storms. In presenting the results, rather than defining a threshold value, the raw indicator values are compared. A probabilistic representation of the indicator is preferred for an ATFM application allowing the user to evaluate the risks in making a decision.

In Fig. 4 results are compared between the NN and baseline indicators for the 7 days in the test data set. From the figure it is evident that the NN model outperforms the baseline indicator given the greater value of AUC. Moreover, because the NN curve is always above the baseline curve, the NN model outperforms the baseline independently of a chosen threshold value. It is important to note that the AUC value is dependent on the particular data set being analyzed. The NN model is good at identifying areas without convection (true negatives), thus analysis of days with few convective storms will yield greater AUC values.

In Fig. 5 results are presented for the entire test data set using the prediction score by class. Histograms are provided for the baseline and neural network model. In the graphs the convection class is shown in red, while the non-convective class is shown in gray. Given the class imbalance in the test data set, the distributions have been normalized so that the two classes occupy the same area in the graphs. Ideally, we would like the two distribution completely separated, with the non-convective (gray) distribution closer to a prediction score 0 and the convective distribution (red) closer to a prediction score of 1. The histogram on the left, shows the baseline model does a good job at evaluating the non-convective areas with a low probability score. However it is also unable to distinguish a large portion of the convective areas from non-convective areas. The histogram on the right, corresponding to the NN model shows less overlap between the two class distributions indicating better performance.

Fig. 6 shows a map representation of the target function alongside the baseline and neural network model predictions for a geographical domain centered over Italy. ROC AUC values corresponding to the data portrayed on the maps are provided. From the figure we can see that while the baseline correctly identifies some areas where storms will develop, it tends to provide low prediction scores and there is a large portion of the convective areas that it misses completely. The NN model although may tend to slightly overestimate the storms, the prediction probability seems to be more gradual for the convective areas. A traffic manager wanting reroute traffic flows around convective weather based on the predictive score from the indicators, would get a more accurate representation of the convective regions in the airspace by using the neural network based convective indicator.

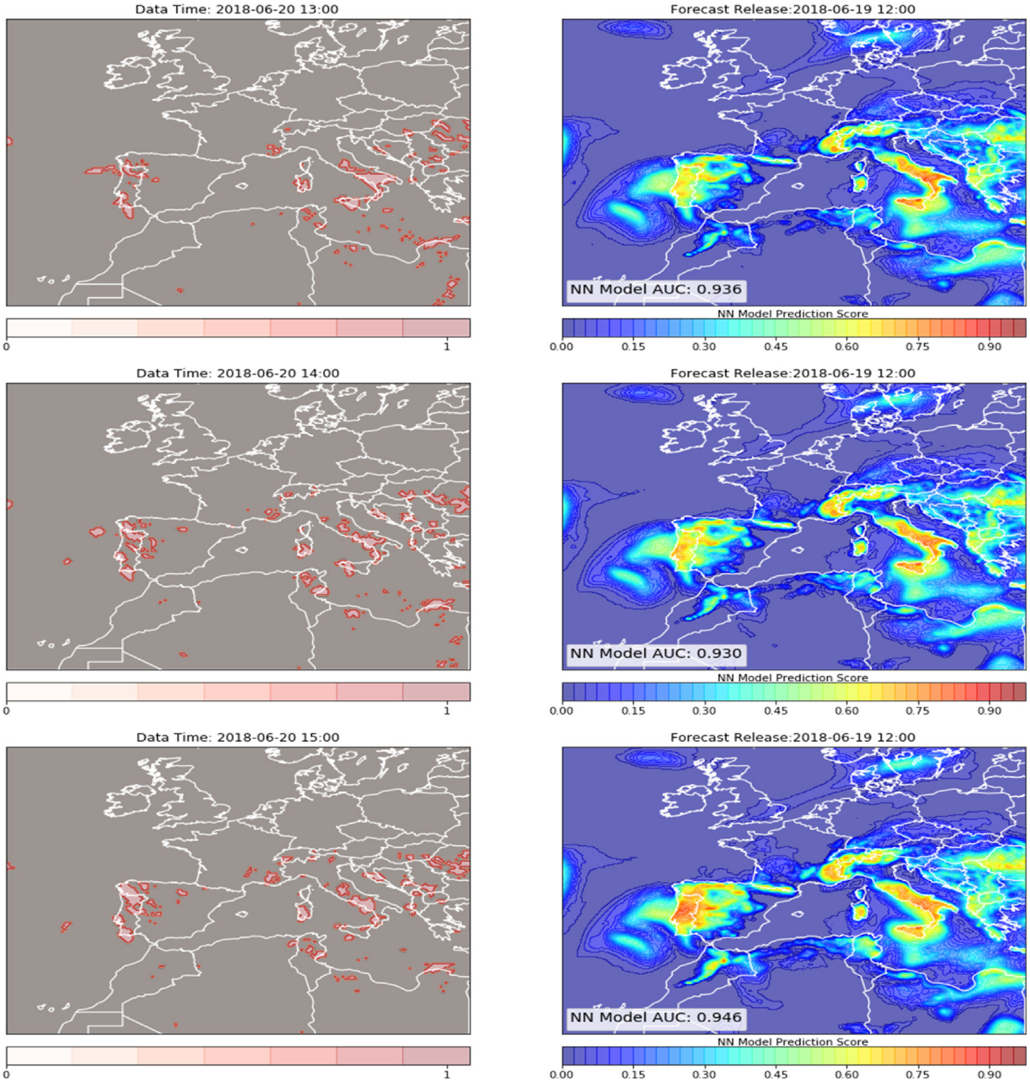
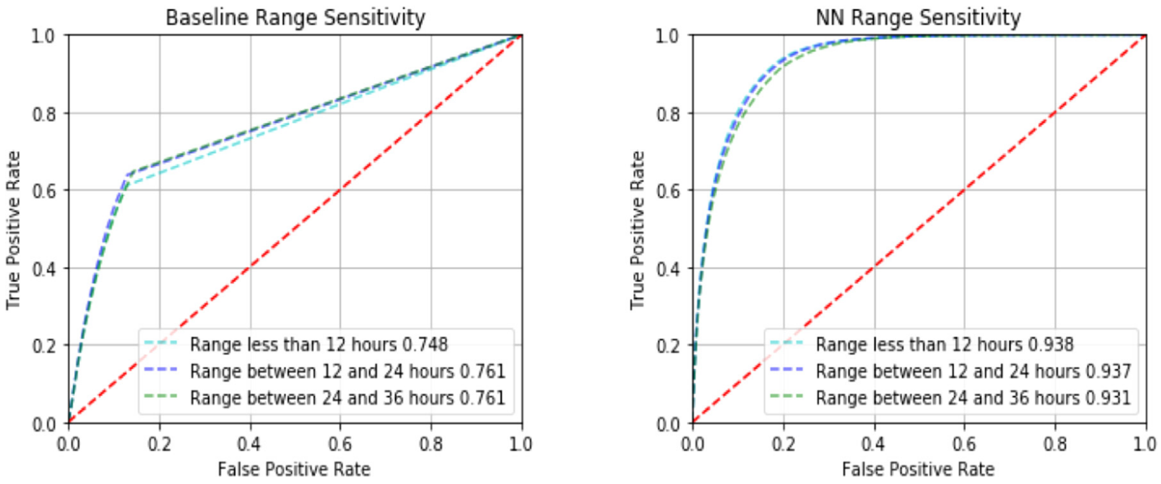

Data Time: 2018-06-20 13:00
0
1
Forecast Release:2018-06-19 12:00
NN Model AUC: 0.936
NN Model Prediction Score
0.00
0.15
0.30
0.45
0.60
0.75
0.90
Data Time: 2018-06-20 14:00
0
1
Forecast Release:2018-06-19 12:00
NN Model AUC: 0.930
NN Model Prediction Score
0.00
0.15
0.30
0.45
0.60
0.75
0.90
Data Time: 2018-06-20 15:00
0
1
Forecast Release:2018-06-19 12:00
NN Model AUC: 0.946
NN Model Prediction Score
0.00
0.15
0.30
0.45
0.60
0.75
0.90

Fig. 7. Convection prediction for June 20, 2018 made on June 19, 12:00.



(a) Baseline Model ROC

(b) NN Model ROC

Fig. 8. ROC curves showing model sensitivity to forecast range variation.

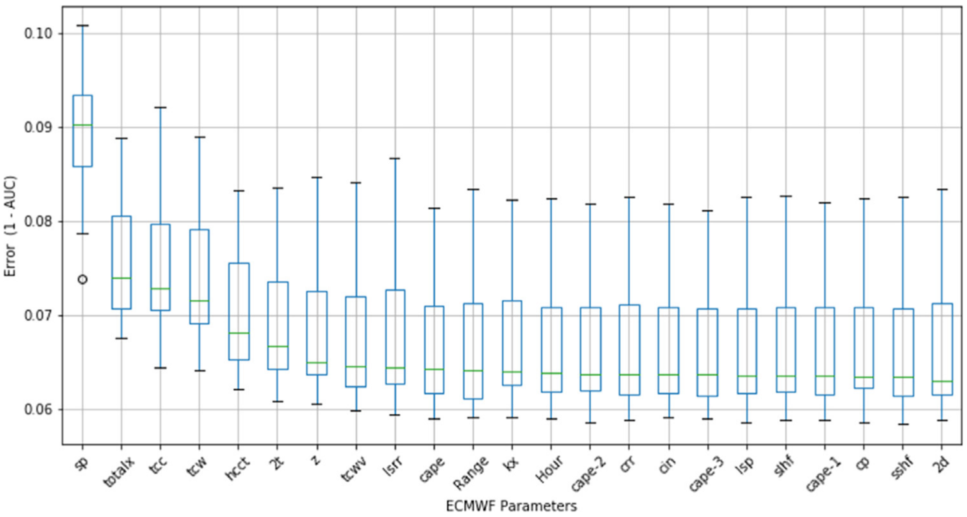
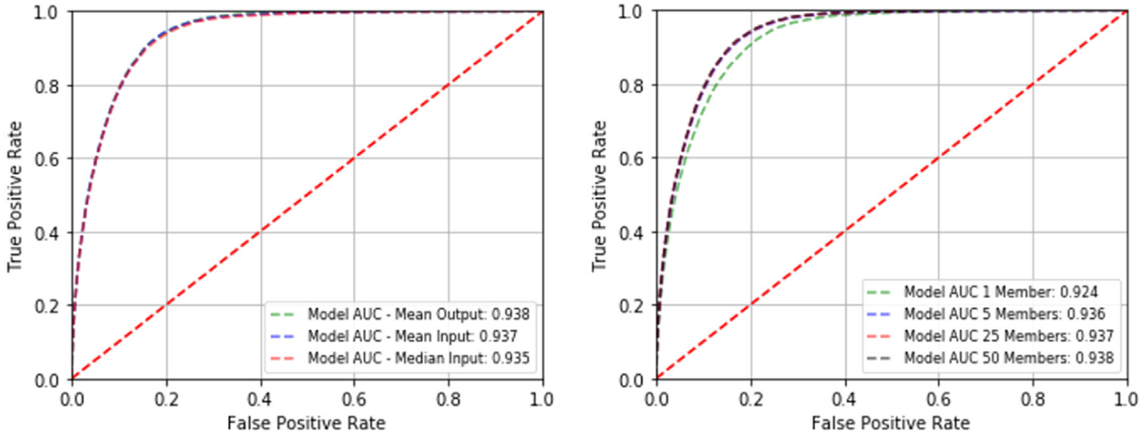


Fig. 9. Error in AUC after permutation of surface ECMWF parameters.



(a) Comparison of ensemble aggregation methods for model evaluation.

(b) NN model results using a subset of the ensemble members.

Fig. 10. ROC curves showing model sensitivity to ensemble data.

In Fig. 7 results for the NN convection indicator are shown for the entire geographical domain. The figure shows the target function and model predictions for June 20, 2018 from 13:00 to 16:00 UTC based on the forecast from June 19, 12:00.

The map representation of results from Figs. 6 and 7 show the convective predictions made on the day before operations. Continuous monitoring of an upcoming convective situations is necessary for an ATFM operations, therefore it is important to understand how the predictions of the indicator change over the prediction time horizon. In Fig. 8, we show how the ROC curves for both indicators behave given different forecast ranges. From the figure can see that the AUC for the NN model remains fairly constant at ranges up to 24 h, and degrades slightly when extended to 36 h. These results indicate that the quality of the results do not degrade at the time scales required for the pre-tactical phase of ATFM.

4.3. Study of feature relevance

A permutation analysis was performed to understand which of the ECMWF EPS parameters were most important in predicting convection. The theory behind the permutation analysis is to measure the importance of a feature by calculating the degradation of model performance after permuting the feature. A more important feature will increase the model error after shuffling its values because the model relies on this feature to make its prediction, while shuffling the values of a less important feature will have little impact on the model error. This technique was first introduced specifically for random forest models (Breiman, 2001), and later expanded to a model-agnostic version (Fisher et al., 2019).

In our analysis we measure the model error by the increase in 1 - AUC. Fig. 9 shows the results of a permutation analysis performed on several batches from the test data set. For each parameter we are able to see the distribution of error associated with permuting that feature. From the figure we see can see that permuting the surface pressure (sp), total totals index (totalx), total cloud cover (tcc) and total column water (tcw) parameters produce the greatest error. Interestingly enough we can relate these parameters to the already mentioned conditions that are favorable to thunderstorms; moisture (tcc, tcw), instability (totalx), and lifting force (sp). This type of analysis will be useful in selecting additional NWP to include in future versions of out model.

4.4. Study of model sensitivity to ensemble data

In this section a series of case studies are presented to better understand the model sensitivity to the ensemble data. A random subset of the test data set is selected to evaluate the model for various cases.

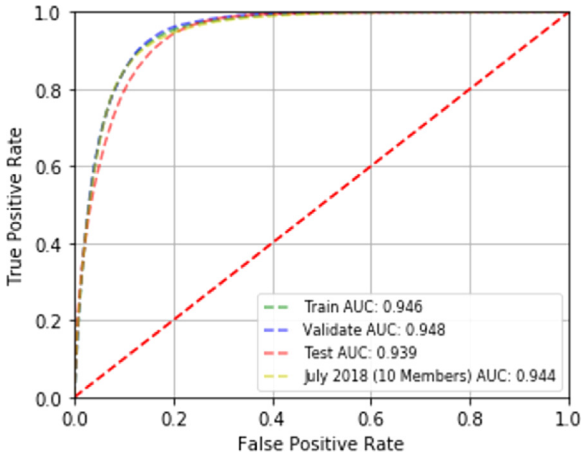


Fig. 11. ROC curve comparing NN model evaluation using subsets of training, validation and test data sets. An additional data set based on forecast predictions for July 2018 is also compared.

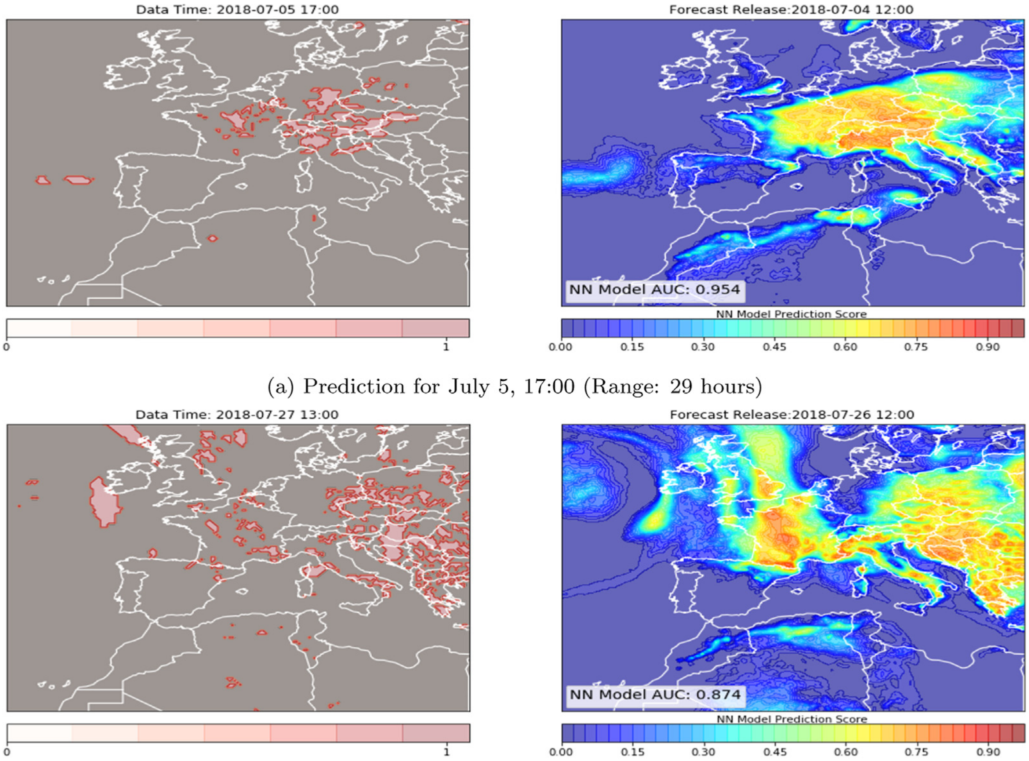
In the first case, various methods of aggregating the ensemble data are compared. Evaluating the model on each individual member and averaging the outputs is compared with taking the mean of each parameter prior to evaluating the model. Additionally, results are also shown for using an input based on the median value for each parameter. In Fig. 10(a) it is shown that while the various ensemble aggregation methods do not impact the results, averaging the output is slightly better.

In the second case, we explore how model results compare if a subset of the ensemble members are used to make a prediction. For this study the same random subset of the test data is evaluated multiple times using a limited number of ensemble members. Results are shown when the model is evaluated using only 1, 5, and 25 randomly selected members and compared with results when using the entire ensemble. From Fig. 10(b), it is evident that results improve as the number of ensemble members used is increased, although the incremental improvement is diminished as more members are added.

In the last case study, the model performance is compared on four data sets; the training, validation and test data partitions, as well as an additional data set comprised of ECMWF predictions for the month of July 2018 using only 10 ensemble members. Within each of these four data groups, 50 randomly selected hourly predictions we used to evaluate the model. A ROC curve comparing the model performance across the four data sets is provided in Fig. 11, from the figure we can see the model classification skill is similar for all data sets. It is important to highlight the performance for the July data set comprised of only 10 ensemble members is similar to that of the other data sets which comprised of all 50 members, this further confirms the results presented in Fig. 10(b). Finally, given that the ROC curve AUC value is sensitive to the weather conditions within each data set, Fig. 12 shows map representations of the July data.

5. ATFM application

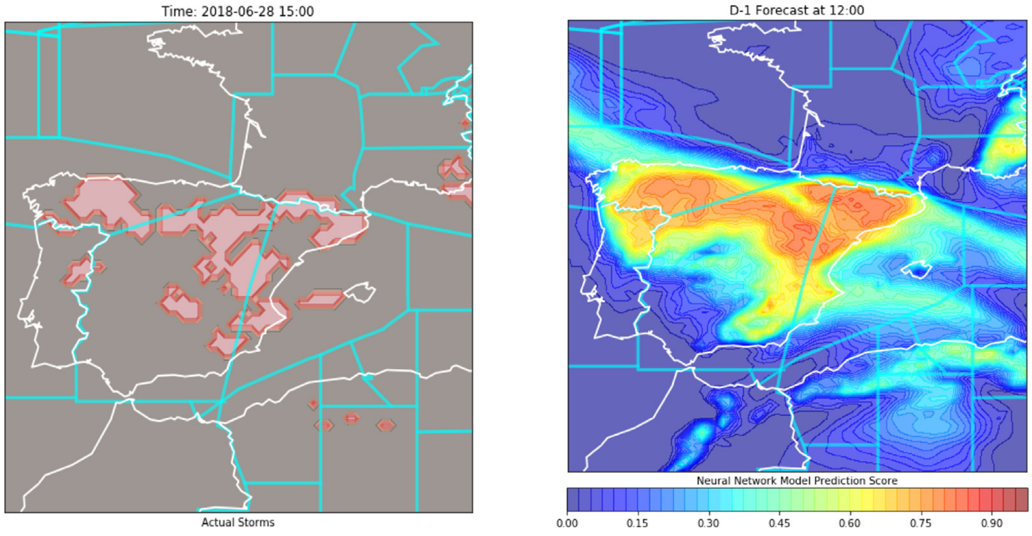
In this section we present an example of possible application of the neural network indicator in an ATFM operational setting. The objective of this work is to provide traffic managers with awareness of where and when convective weather will develop. Perhaps, the most obvious application would be to overlay the convective prediction on a map of structured airspace, in this way traffic managers could have information on which sectors will be impacted be convective weather. A conceptual map of our indicator overlaid atop the European Area Control Centres (ACC), ACCs establish the areas of jurisdiction for the various control units in the European airspace. In Fig. 13 we compare the actual storm situation as captured by the RDT data with the convection prediction of the NN indicator. From the figure we can


Data Time: 2018-07-05 17:00
0
1
Forecast Release:2018-07-04 12:00
NN Model AUC: 0.954
NN Model Prediction Score
0.00
0.15
0.30
0.45
0.60
0.75
0.90
Data Time: 2018-07-27 13:00
0
1
Forecast Release:2018-07-26 12:00
NN Model AUC: 0.874
NN Model Prediction Score
0.00
0.15
0.30
0.45
0.60
0.75
0.90

(a) Prediction for July 5, 17:00 (Range: 29 hours)

(b) Prediction for July 27, 13:00 (Range: 25 hours)

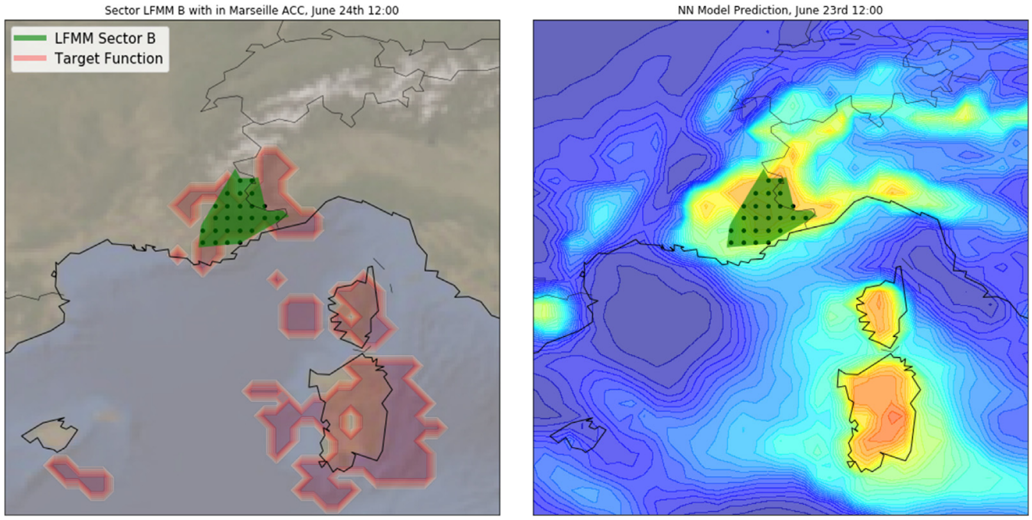
Fig. 12. Convection predictions for July 2018 based on 10 members from ensemble.



(a) Actual Storm Data

(b) Neural network model prediction

Fig. 13. Convection prediction captures storms in Spanish ACCs one day before.


Sector LFMM B with in Marseille ACC, June 24th 12:00
LFMM Sector B
Target Function
NN Model Prediction, June 23rd 12:00

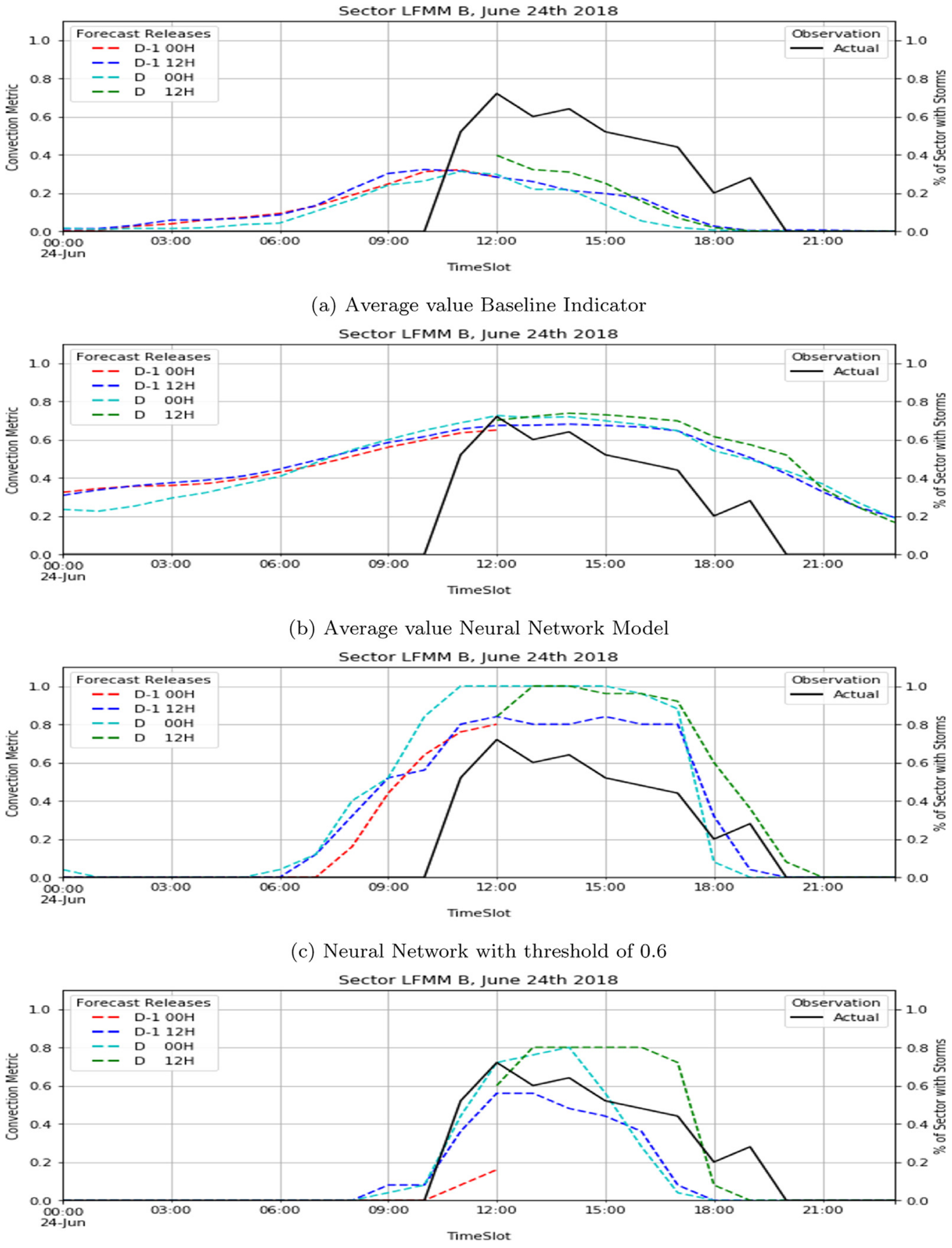
(a) Target function representation

(b) Nueral network model prediction

Fig. 14. Marseille Sector B and convective weather situation on June 24th, 2018 at 12:00 and prediction (Range: 24 h.).

see that there was storm activity in multiple Spanish ACCs on June 28th, 2018 at 15:00 UTC, the neural network indicator prediction one day before the day of operations (D-1) at noon is able to capture the general area of the storms. In another application the information is presented in a manner that is specific for a unit of airspace. In this example we focus on he Marseille ACC, a region of airspace responsible for 15.2% of ATFM delay in Europe in 2018 (EUROCONTROL, 2019). Specifically we focus on Sector B within the Marseille ACC as shown in Fig. 14. Based on the NWP resolution, the area covering this unit of airspace can be represented with 25 grid points. Using the model predictions from these 25 points it is possible to define a metric to evaluate the convection situation in the sector. In Fig. 15 multiple convection metrics based on the baseline and NN model are compared with the RDT data from June 24th 2018. Figs. 15(a) and 15(b) show metrics based on the average indicator value over the 25 points for the baseline and NN model. In Figs. 15(c) and 15(d) the metric is based on the NN model output and the percentage of grid points exceeding specific thresholds. The various dashed colored lines corresponding to the left y-axis relate to the calculated convection metric with the Marseille ACC for various forecast releases on the day before operations (D-1) and the day of operations (D). The solid black line corresponding to the right *y*-axis, shows the percentage of airspace region with storms according to the target function.

From Fig. 15 it is evident that while all metrics capture some convective activity, using the neural network model results with an applied threshold better captures the convective situation within LFMM Sector B. It is imagined that the neural network model output can be used to define convection metrics within European airspace to continuously monitor and assess the weather situation. Further analysis is needed to better understand how these convection metrics impact ATFM attributes such as airspace capacities, traffic demand, and weather regulations. Understanding the relationship between weather prediction and the impact on the traffic would allow traffic managers to make better decisions during the pre-tactical phase of ATFM.



(a) Average value Baseline Indicator

(b) Average value Neural Network Model

(c) Neural Network with threshold of 0.6

(d) Neural Network with threshold of 0.7

Fig. 15. Convection metrics evaluating the weather situation in Marseille Sector B for multiple forecast releases.

6. Summary and conclusions

In this paper we have applied machine learning techniques to predict convective areas within the next 36 h. By combining data from satellite storm observations and ensemble NWP products, a neural network algorithm is trained to predict the occurrence of convective weather. The NN model is able to outperform an existing convection indicator currently used in aviation applications. Analyses of the model on a test data set indicate that model performance does not degrade significantly for forecast ranges up to 36 h. Additional evaluation of the model showed that model performance is maintained when evaluating on a subset of the ensemble forecast. Furthermore, a permutation analysis was completed to detect which EPS parameters are most relevant to convection prediction. Findings confirm those parameters related to the physical process of convection, correspond to the most relevant features of our model. Lastly, examples are provided for the use of the indicator in an ATFM operational setting. Visualization of model predictions show that the model is able to accurately predict regions where convection will develop. Model predictions are used to develop convection metrics and used to evaluate the weather situating within a specific sector within the Marsielle ACC and compared against storm observations. This analysis suggest that applying a threshold atop of the model predictions can improve the detection of convective weather.

Despite these initial positive results, several areas of improvement remain to be tackled in future efforts. One area of improvement is to move away from the assumptions of treating each ensemble member and each grid point as independent. It is acknowledged that more efficient use of the NWP ensemble product would be to provide model input that jointly considers all ensemble members. Additional data processing and integration of the NWP data is required to provide the model with an input representative of all ensemble members. Furthermore, other model architectures including Convolution Neural Networks and Long Short-Term Memory Networks need to be considered to better extract the spatial–temporal relationships within the data.

Another area of improvement is the quality of the data that is used to train the models. Making use of higher resolution NWP products as well as additional parameters at various atmospheric levels could provide improved model inputs. Additionally, we could also incorporate other sources of convection observation data, such as radar or lightning, to provide the model with a more precise target function to be used during training. Future research efforts should focus on how to best integrate these various data sources.

Furthermore, the model uses a binary classification scheme to predict the probability that convection will occur. However, in the future we hope to expand the model to also identify key characteristics associated with convection, such as storm severity and cloud top altitude; both relevant information in an air traffic flow management context. These efforts could be accomplished by moving away from a binary representation and elaborating a more sophisticated target function able to capture those storm characteristics that have a major impact on ATFM operations.

Lastly, an important step in the application of the model in an ATFM operation setting is further refinement of raw model output in order to provide traffic managers with simple and relevant information. One possible solution is to translate the model output into a color-scheme, similar to what is currently in use today.

The goal of this research is to provide traffic managers with improved convective weather information at time frames compatible with pre-tactical ATFM planning. While this objective has been achieved to some extent, further research efforts are still needed to relate the convection prediction with ATFM metrics such as airspace capacities, traffic demand, and ATFM mitigation strategies. Only then can the full benefit to ATFM operations be achieved.

CRediT authorship contribution statement

Aniel Jardines: Conceptualization, Methodology, Data curation, Writing - original draft. **Manuel Soler:** Supervision, Conceptualization, Funding acquisition, Writing - review & editing. **Alejandro Cervantes:** Supervision, Methodology, Writing - review & editing. **Javier García-Heras:** Supervision, Writing - review & editing. **Juan Simarro:** Conceptualization, Resources, Review.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would like to acknowledge the financial support of the Spanish Ministry of Science, Innovation and Universities under grant RTI2018-098471-B-C32. The project has received funding from the SESAR Joint Undertaking within the framework of the European Union's Horizon 2020 research and innovation programme under grant agreement No 891965.

References

Baldauf, M., Seifert, A., Förstner, J., Majewski, D., Raschendorfer, M., & Reinhardt, T. (2011). Operational convective-scale numerical weather prediction with the COSMO model: Description and sensitivities. *Monthly Weather Review*, *139*(12), 3887–3905.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Collins, W., & Tissot, P. (2015). An artificial neural network model to predict thunderstorms within 400 km2 south texas domains. *Meteorological Applications*, *22*(3), 650–665.

Cook, A., & Tanner, G. (2015). European airline delay cost reference values. *Eurocontrol: Brussels, Belgium*.

EUROCONTROL (2019). *2018 performance review report, "An assessment of air traffic management in Europe during the calendar year 2018"*. EUROCONTROL, https://www.eurocontrol.int/air-navigation-services-performance-review.

Evans, J. E., & Ducot, E. R. (2006). Corridor integrated weather system. *Lincoln Laboratory Journal*, *16*(1), 59.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81.

González-Arribas, D., Soler, M., Sanjurjo-Rivo, M., García-Heras, J., Sacher, D., Gelhardt, U., Lang, J., Hauf, T., & Simarro, J. (2019). Robust optimal trajectory planning under uncertain winds and convective risk. In E. N. R. Institute (Ed.), *Air traffic management and systems III* (pp. 82–103). Singapore: Springer Singapore.

He, J., & Loboda, T. V. (2020). Modeling cloud-to-ground lightning probability in alaskan tundra through the integration of weather research and forecast (WRF) model and machine learning method. *Environmental Research Letters*, *15*(11), Article 115009.

Khandan, R., Alavipanah, S. K., Biazar, A. P., & Gharaylou, M. (2018). Probabilistic convective initiation nowcasting with reduced satellite-NWP predictors over Iran. *Asia-Pacific Journal of Atmospheric Sciences*, *54*(3), 431–443.

Lee, J.-G., Min, K.-H., Park, H., Kim, Y., Chung, C.-Y., & Chang, E.-C. (2020). Improvement of the rapid-development thunderstorm (RDT) algorithm for use with the GK2a satellite. *Asia-Pacific Journal of Atmospheric Sciences*, *56*(2), 307–319.

Li, H., Li, Y., Li, X., Ye, Y., Li, X., & Xie, P. (2019). A comparative study on machine learning approaches to thunderstorm gale identification. In *Proceedings of the 2019 11th international conference on machine learning and computing* (pp. 12–16). New York, NY, USA: Association for Computing Machinery.

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316.

Mecikalski, J. R., Williams, J. K., Jewett, C. P., Ahijevych, D., LeRoy, A., & Walker, J. R. (2015). Probabilistic 0–1-h convective initiation nowcasts that combine geostationary satellite observations and numerical weather prediction model data. *Journal of Applied Meteorology and Climatology*, *54*(5), 1039–1059.

Molteni, F., Buizza, R., Palmer, T. N., & Petroliagis, T. (1996). The ECMWF ensemble prediction system: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, *122*(529), 73–119.

Oxf (2015). *ATPL ground training, CAE Oxford aviation academy METEOROLOGY*. CAE Oxford Aviation Academy (UK).

Palmer, T., Buizza, R., Hagedorn, R., Lawrence, A., Leutbecher, M., & Smith, L. (2006). Ensemble prediction: a pedagogical perspective. *ECMWF Newsletter*, *106*(106), 10–17.

Simon, T., Fabsic, P., Mayr, G. J., Umlauf, N., & Zeileis, A. (2018). Probabilistic forecasting of thunderstorms in the Eastern Alps. *Monthly Weather Review*, *146*(9), 2999–3009.

Spiridonov, V., Baez, J., Telenta, B., & Jakimovski, B. (2020). Prediction of extreme convective rainfall intensities using a free-running 3-D sub-km-scale cloud model initialized from WRF km-scale NWP forecasts. *Journal of Atmospheric and Solar-Terrestrial Physics*, *209*, Article 105401.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Šaur, D. (2017). Forecasting of convective precipitation through NWP models and algorithm of storms prediction. In R. Silhavy, R. Senkerik, Z. Kominkova Oplatkova, Z. Prokopova, & P. Silhavy (Eds.), *Artificial intelligence trends in intelligent systems* (pp. 125–136). Cham: Springer International Publishing.

Wilson, J. W., Crook, N. A., Mueller, C. K., Sun, J., & Dixon, M. (1998). Nowcasting thunderstorms: A status report. *Bulletin of the American Meteorological Society*, *79*(10), 2079–2100.

Zhou, K., Zheng, Y., Li, B., Dong, W., & Zhang, X. (2019). Forecasting different types of convective weather: A deep learning approach. *Journal of Meteorological Research*, *33*(5), 797–809.