Use of Machine learning for the detection and classification of observation anomalies

Mohamed Dahoui

ECMWF, Shinfield Park, Reading, UK

mohamed.dahoui@ecmwf.int

Introduction

For the last few years, an automatic data checking system has been used at ECMWF to monitor the quality and availability of observations processed by ECMWF's data assimilation system (Dahoui *et al.*, 2020). The tool is playing an important role in flagging up observation issues and enabling timely triggering of mitigating actions. The system has few weakness:

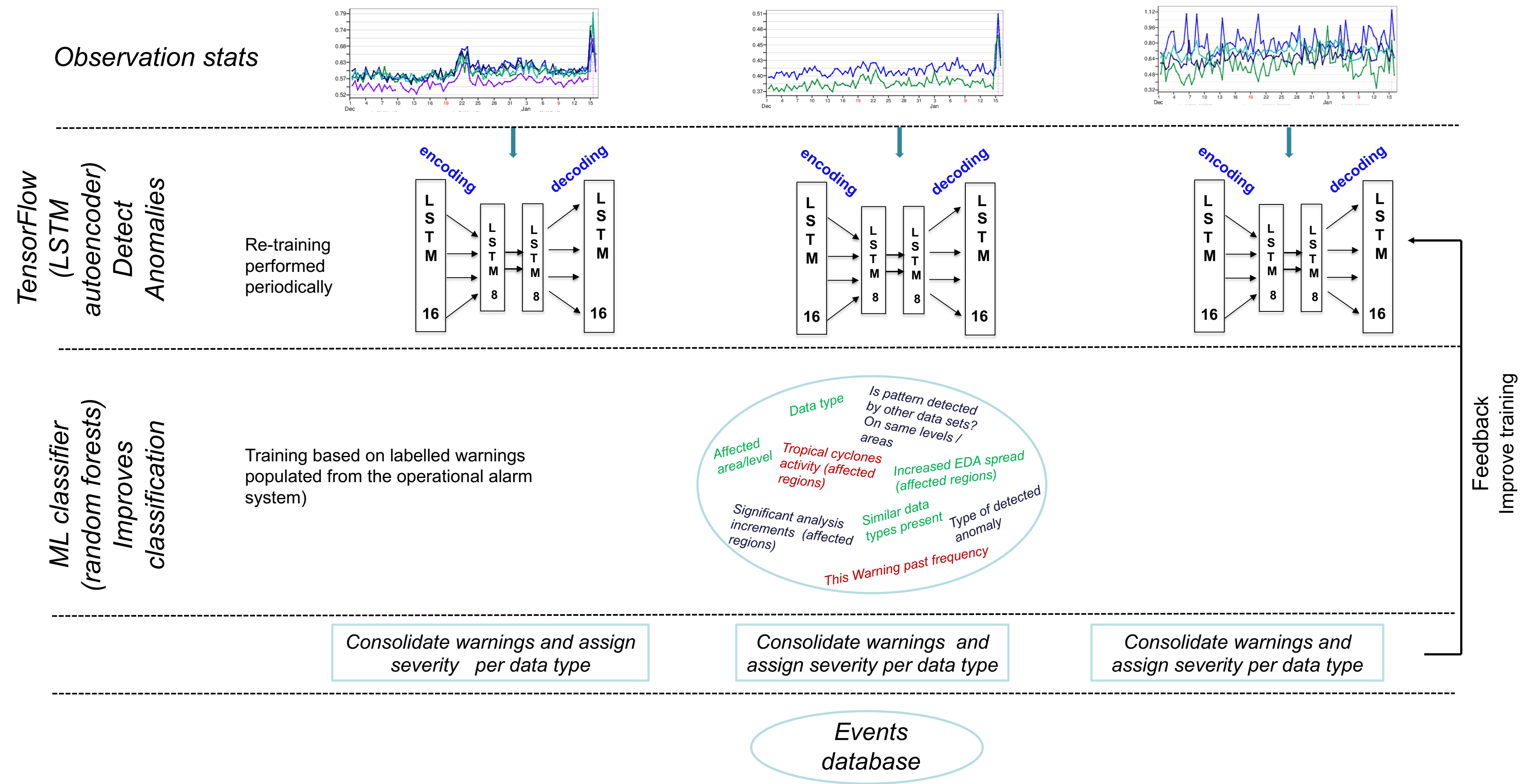
- The behaviour of the system is less optimal when assigning a severity level to detected events. occasionally less significant events can be communicated as severe. When the day-to-day variability is small, moderate changes can be interpreted as severe from a statistical point of view. Not every threshold violation is a problem,
- The current system is not able to consider warnings affecting individual data types in the context of what is happening with the rest of the observing system and the type of weather activity dominating the affected areas. Most anomaly detection tests are based on first-guess departures which are the combination of uncertainties from observations and short-range forecast. As a results, generated warnings are not necessarily caused by observation problems. Factors causing the statistics to deviate are diverse.

Machine learning techniques offer the possibility to improve the anomaly detection via a better detection of patterns, and to improve the classification of events by severity and cause. They do not need a periodic adjustment of threshold limits, either, which makes them useful for the monitoring of satellite data from a growing number of satellite platforms.

A new version of the automatic data checking system has been designed. It is based on an unsupervised recurrent neural network algorithm for the detection of abnormal statistics, and on a supervised learning algorithm (random forest) to classify the detected events.

Design of the machine learning observational data checking system

The anomaly detection module rely on an unsupervised neural network algorithm to detect large deviations of statistics. This module aims to flag up sudden changes and slow drifts of statistics. The anomaly detection is performed separately for all observation types. The combined results for all observation types are analysed by a supervised machine learning classifier (random forest) to adjust the severity (including a dismissal of the event), indicate the likely cause, and suggest whether action is needed. The classification results are then processed for each individual data type in order to generate relevant plots and archive warnings in an event database.



Observation stats
TensorFlow (LSTM autoencoder) Detect Anomalies
Re-training performed periodically
encoding
decoding
LSTM 16
LSTM 8
LSTM 8
LSTM 16
ML classifier (random forests) Improves classification
Training based on labelled warnings populated from the operational alarm system)
Data type
Is pattern detected by other data sets? On same levels / areas
Affected area/level
Tropical cyclones activity (affected regions)
Increased EDA spread (affected regions)
Significant analysis increments (affected regions)
Similar data types present
Type of detected anomaly
This Warning past frequency
Consolidate warnings and assign severity per data type
Consolidate warnings and assign severity per data type
Consolidate warnings and assign severity per data type
Events database
Feedback
Improve training

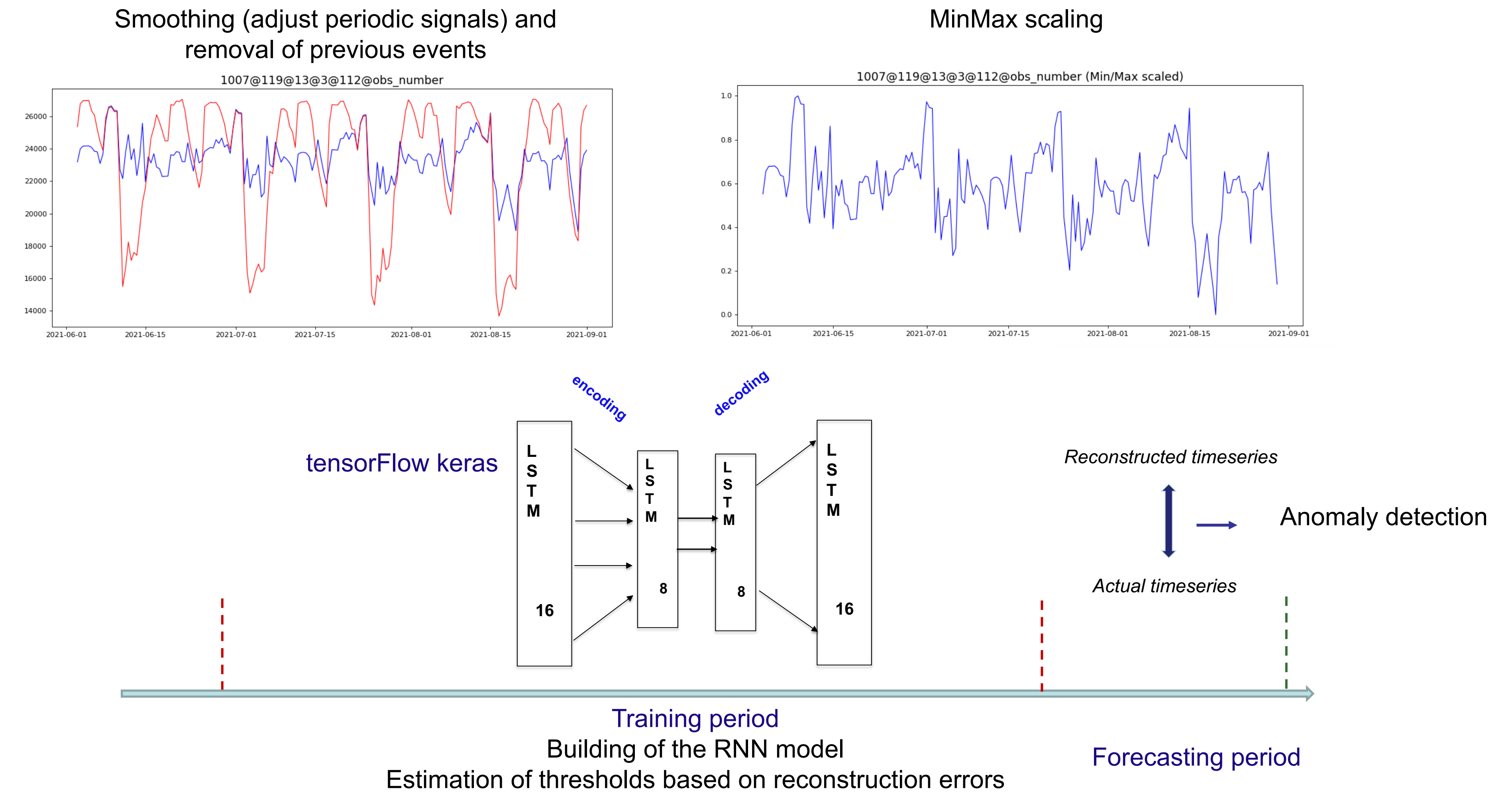
Figure 1: Schematic of the data checking system. The autoencoder LSTM has five layers. The first two encoding layers (with 16 and 8 units respectively) are designed to create a compressed representation of the input data. The third layer processes the compressed vector to provide input for the subsequent decoding layers, and the last two decoding layers (with 8 and 16 units respectively) aim to reconstruct the input data from the compressed representation.

Unsupervised detection of observation anomalies

Two neural network models are applied to each individual data group to learn from the short-term behaviour (past three months) and the long-term evolution (past 12 months when available). The neural networks are autoencoders with long short-term memory (LSTM) cells. The choice of LSTM is mainly intended to enable multi-feature analysis, which is important to support large amounts of data. LSTM offers also the possibility to learn the temporal evolution of statistics.

The short-term model is trained every data assimilation cycle using recent statistics and excluding the last two days. The training dataset contains only statistics that are considered to be 'normal'. Previously detected events and outliers are excluded. As part of the training, we determine the resulting reconstruction error, which is conservatively chosen as the upper tail of the calculated loss in the training set. The trained model is then applied to the latest data sample (spanning the last few days) to reconstruct/predict the current statistics. statistics will be larger than the reconstruction error when abnormal statistics are encountered (Figure 2). Statistics that are provided as input to the short-term model are Min/Max normalised. When relevant, the statistics are adjusted to remove periodic signals.

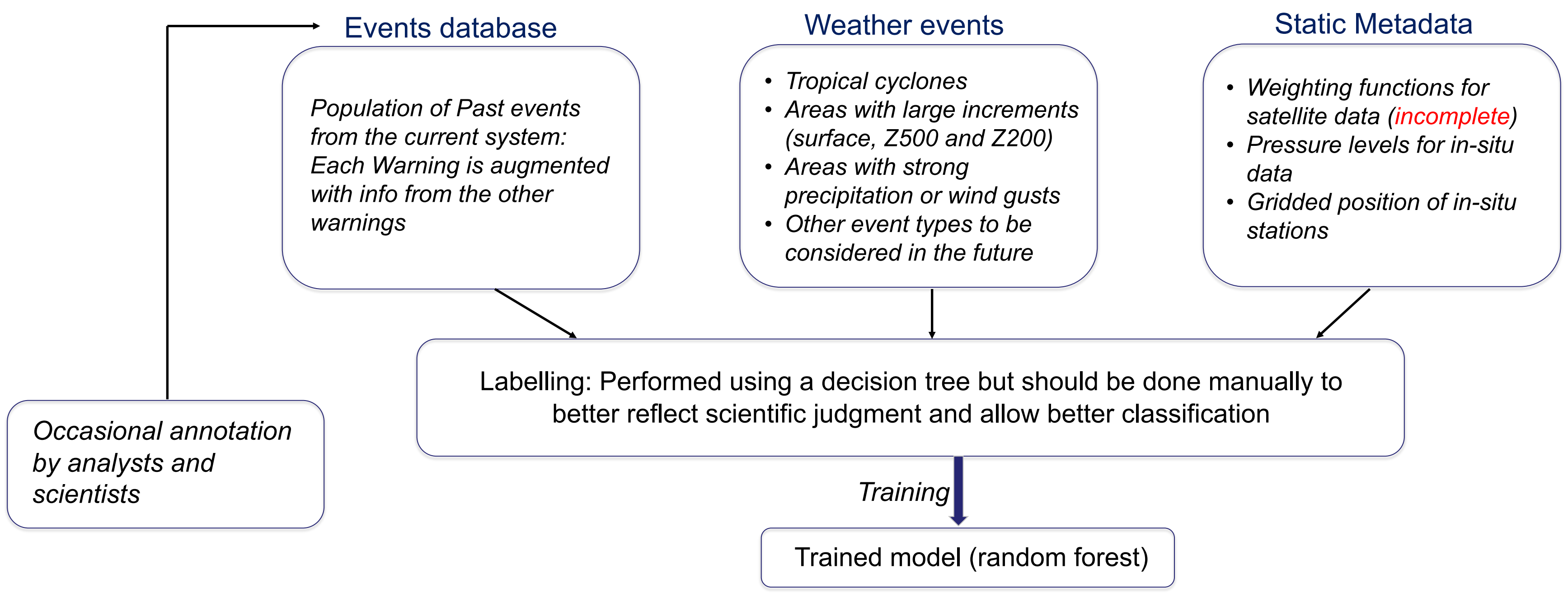
The aim of the long-term trained model is to detect a slow drift of statistics. The model is trained once every quarter using the past 12 months of statistics (if available). To speed up the training and smoothen day-to-day variability, the data are sampled over periods of ten days. As part of the training, we determine the resulting reconstruction error, which is chosen as the upper tail of the calculated loss in the training set. The trained model is then applied to the latest data sample (spanning the last few weeks sampled every10 days) to reconstruct the current statistics. Large differences between reconstructed statistics and observed ones indicate a significant change compared to long-term behaviour. Such a change can take the form of a step change (due to a model upgrade), or a slow drift of the observation quantity being monitored. The main interest is to detect a slow drift of statistics. This is achieved thanks to a monotonic slope detection algorithm applied to cases flagged up by the neural network. If the slope is not monotonic, the event is discarded.



Smoothing (adjust periodic signals) and removal of previous events
1007@119@13@3@112@obs_number
MinMax scaling
1007@119@13@3@112@obs_number (Min/Max scaled)
tensorFlow keras
encoding
decoding
LSTM 16
LSTM 8
LSTM 8
LSTM 16
Reconstructed timeseries
Actual timeseries
Anomaly detection
Training period
Building of the RNN model
Estimation of thresholds based on reconstruction errors
Forecasting period

Figure 2: Schematic of unsupervised anomaly detection

Supervised classification of detected anomalies

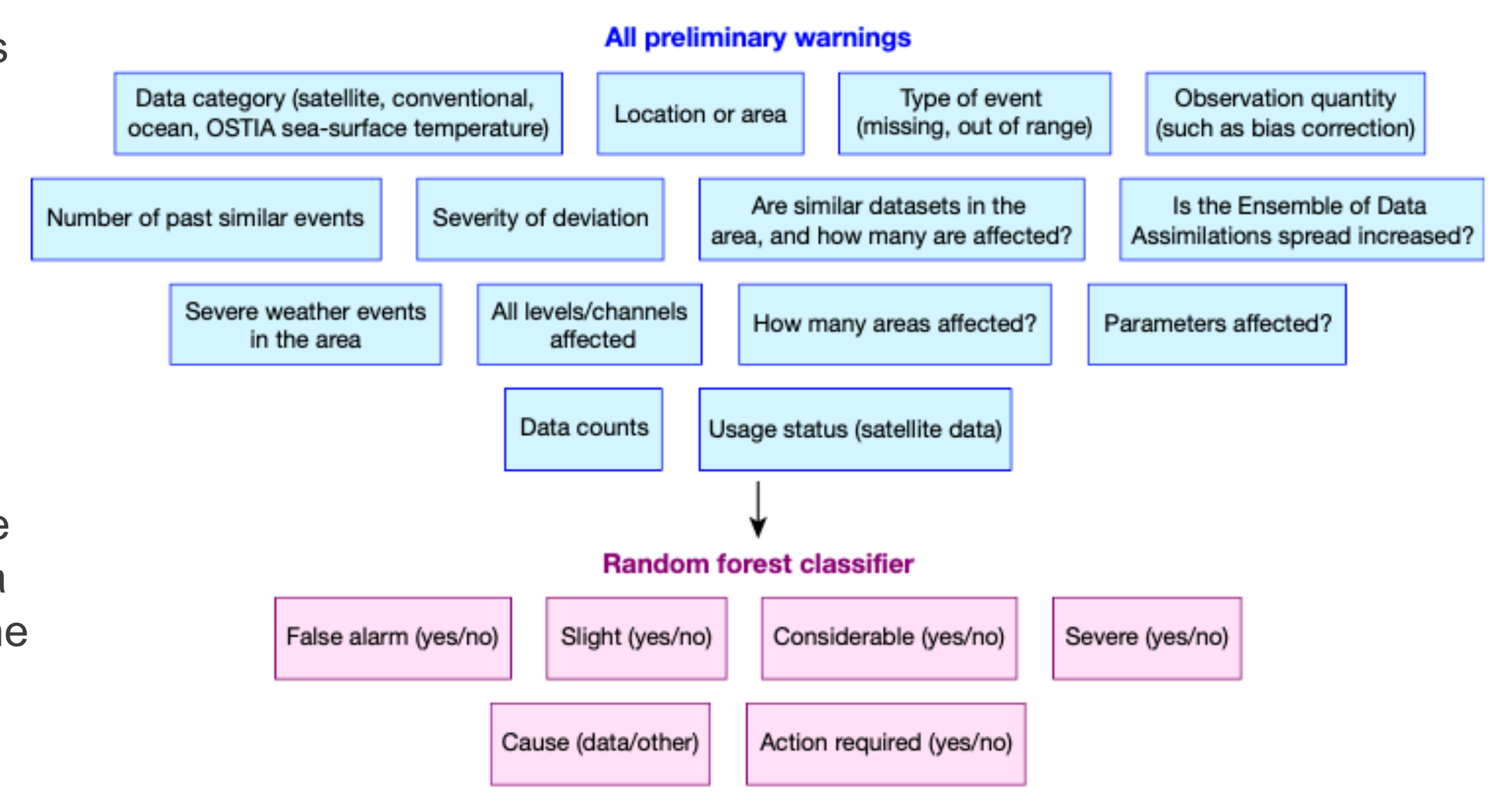


Events database
Population of Past events from the current system: Each Warning is augmented with info from the other warnings
Weather events
Tropical cyclones
Areas with large increments (surface, Z500 and Z200)
Areas with strong precipitation or wind gusts
Other event types to be considered in the future
Static Metadata
Weighting functions for satellite data (incomplete)
Pressure levels for in-situ data
Gridded position of in-situ stations
Labelling: Performed using a decision tree but should be done manually to better reflect scientific judgment and allow better classification
Occasional annotation by analysts and scientists
Training
Trained model (random forest)

Figure 3: Process to train the random forest classifier.

Once the anomaly detection has been performed separately for all data types, all detected events are grouped together in a warning basket. Each event is then augmented by a list of additional features reflecting common events from other data types, significant weather conditions, and the number of past occurrences of the event. A machine learning classifier (random forest) is then applied to define attributes of the detected warnings. These include false alarm (yes/no), slight event (yes/no), considerable event (yes/no), severe event (yes/no), cause (data/other) and action required (yes/no). The machine learning classifier has been trained using a population of previously generated warnings from the current operational system.

The training set has been labelled to define the target attributes. Through the training process, the system is expected to learn the rules that lead to labelling decisions based on event attributes (see Figure 3).



All preliminary warnings
Data category (satellite, conventional, ocean, OSTIA sea-surface temperature)
Location or area
Type of event (missing, out of range)
Observation quantity (such as bias correction)
Number of past similar events
Severity of deviation
Are similar datasets in the area, and how many are affected?
Is the Ensemble of Data Assimilations spread increased?
Severe weather events in the area
All levels/channels affected
How many areas affected?
Parameters affected?
Data counts
Usage status (satellite data)
Random forest classifier
False alarm (yes/no)
Slight (yes/no)
Considerable (yes/no)
Severe (yes/no)
Cause (data/other)
Action required (yes/no)

Figure 4: Features used in the machine learning classifier

The Hunga-Tonga eruption

The eruption of the Hunga Tonga–Hunga volcano in the southern Pacific Ocean on 15 January 2022 had a strong effect on the signals of satellite instruments used by ECMWF. Infrared and Microwave measurements witnessed a powerful pressure and temperature wave that moved quickly from the ocean surface up into the stratosphere and radiated outward. The eruption triggered numerous alerts in the ECMWF data monitoring system because the observed radiance characteristics deviated from the expected behaviour. The Machine learning based system is able to detect the abnormality of the event and to indicate that the origin is likely not related to data issues.

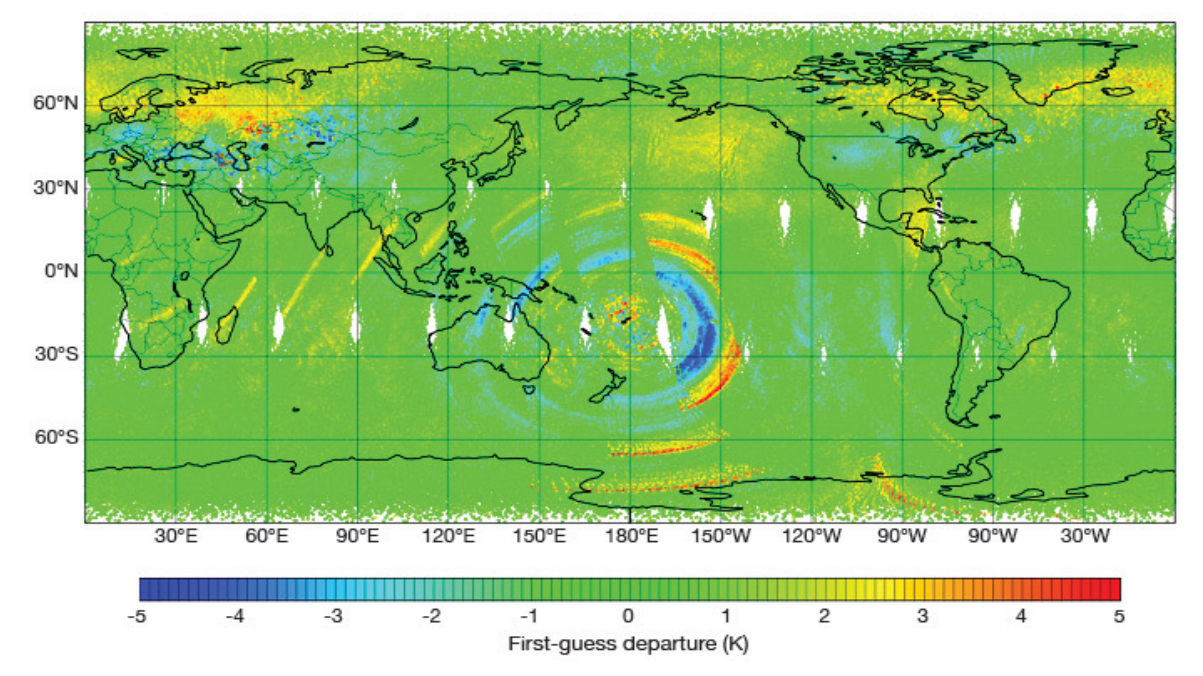


Figure 5: First-guess departures from channel 92 of the IASI infrared interferometer sounder over 24 hours on 15 January 2022.

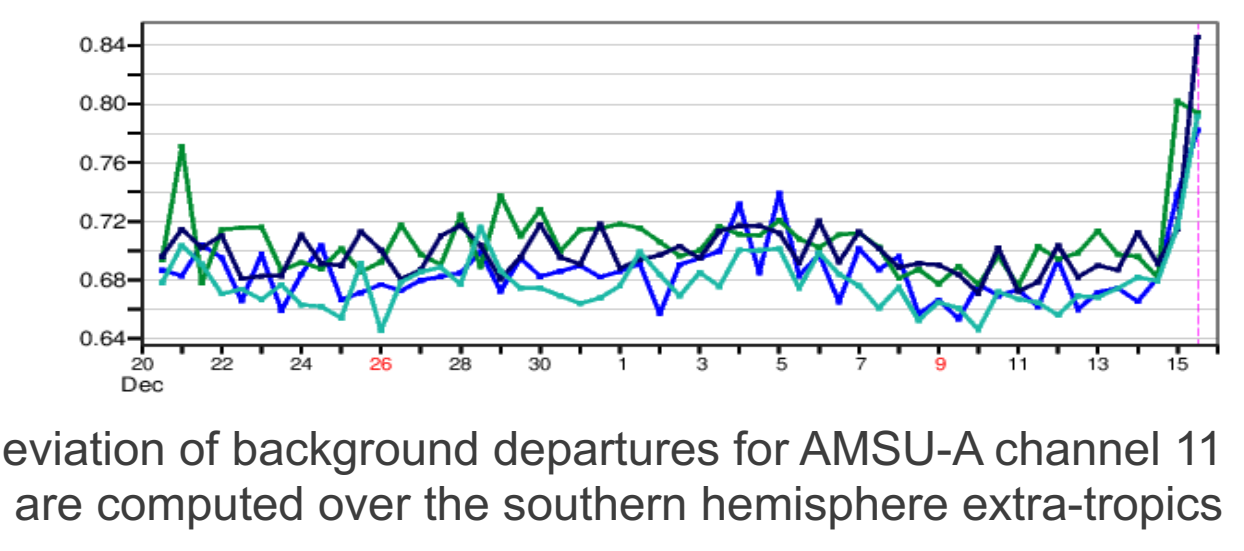


Figure 6: Time series of normalised standard deviation of background departures for AMSU-A channel 11 from four different satellites. The statistics are computed over the southern hemisphere extra-tropics

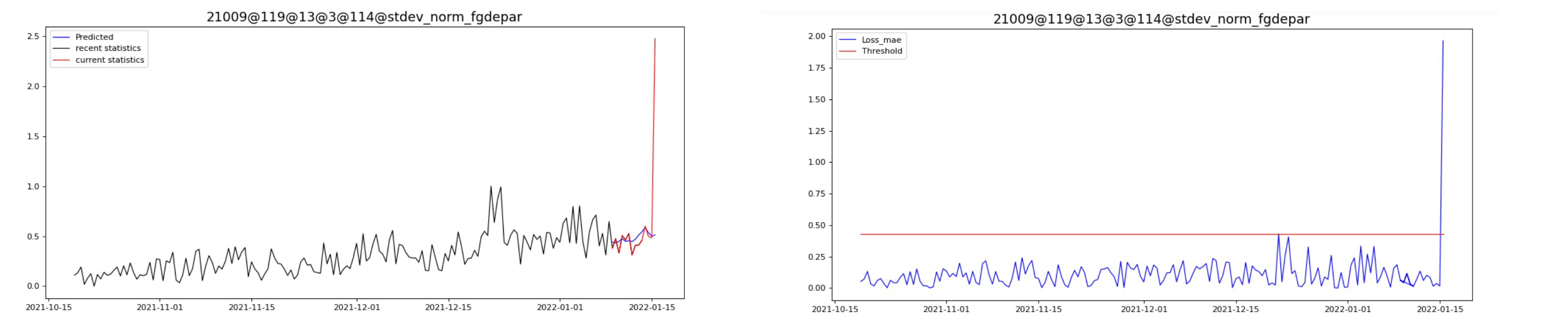


Figure 7: Left panel shows the ML predicted statistics (red) versus recent and current statistics being checked for Channel 13 from Metop-B AMSUA. The right panel shows the prediction error (blue) compared to the threshold

Faulty moored buoy (tropical cyclone FANI)

As the tropical cyclone FANI (April 2019) was progressing in the bay of Bengal a moored buoy was hit by large waves and capsized ahead of the TC. As a result, first guess departures increased considerably. Few of these observations were used by the data assimilation (helped by a relaxed QC due to the increased EDA spread around the TC). The wrong observations degraded significantly the analysis and subsequent forecasts of the TC. The automatic detection system was able to detect these gross errors. With future improvements to the data selection procedure, similar gross errors can be detected and potentially excluded quickly.

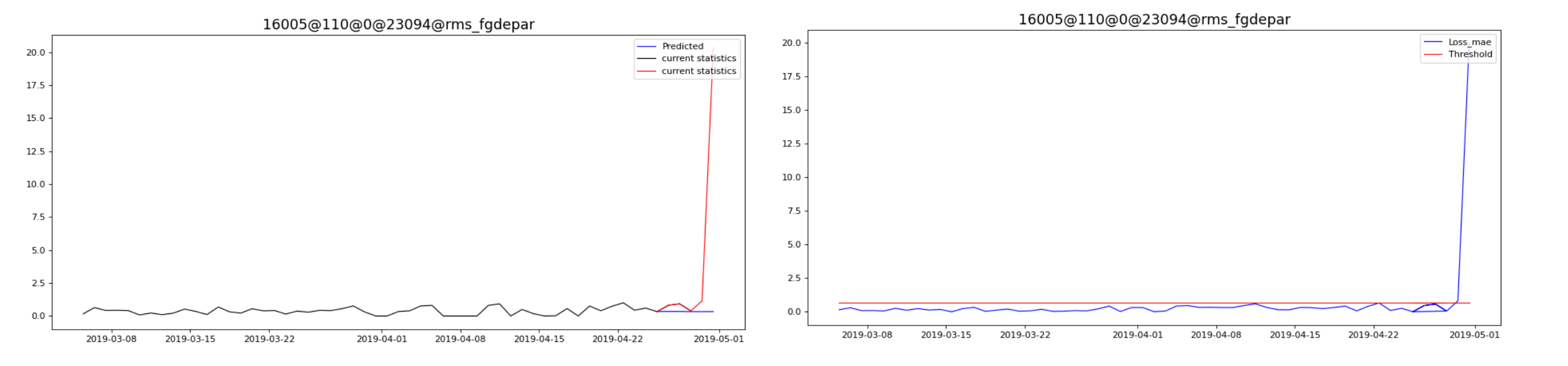


Figure 8: Left panel shows the ML predicted statistics (red) versus recent and current statistics being checked for surface pressure from moored buoy 23094 . The right panel shows the prediction error (blue) compared to the threshold.

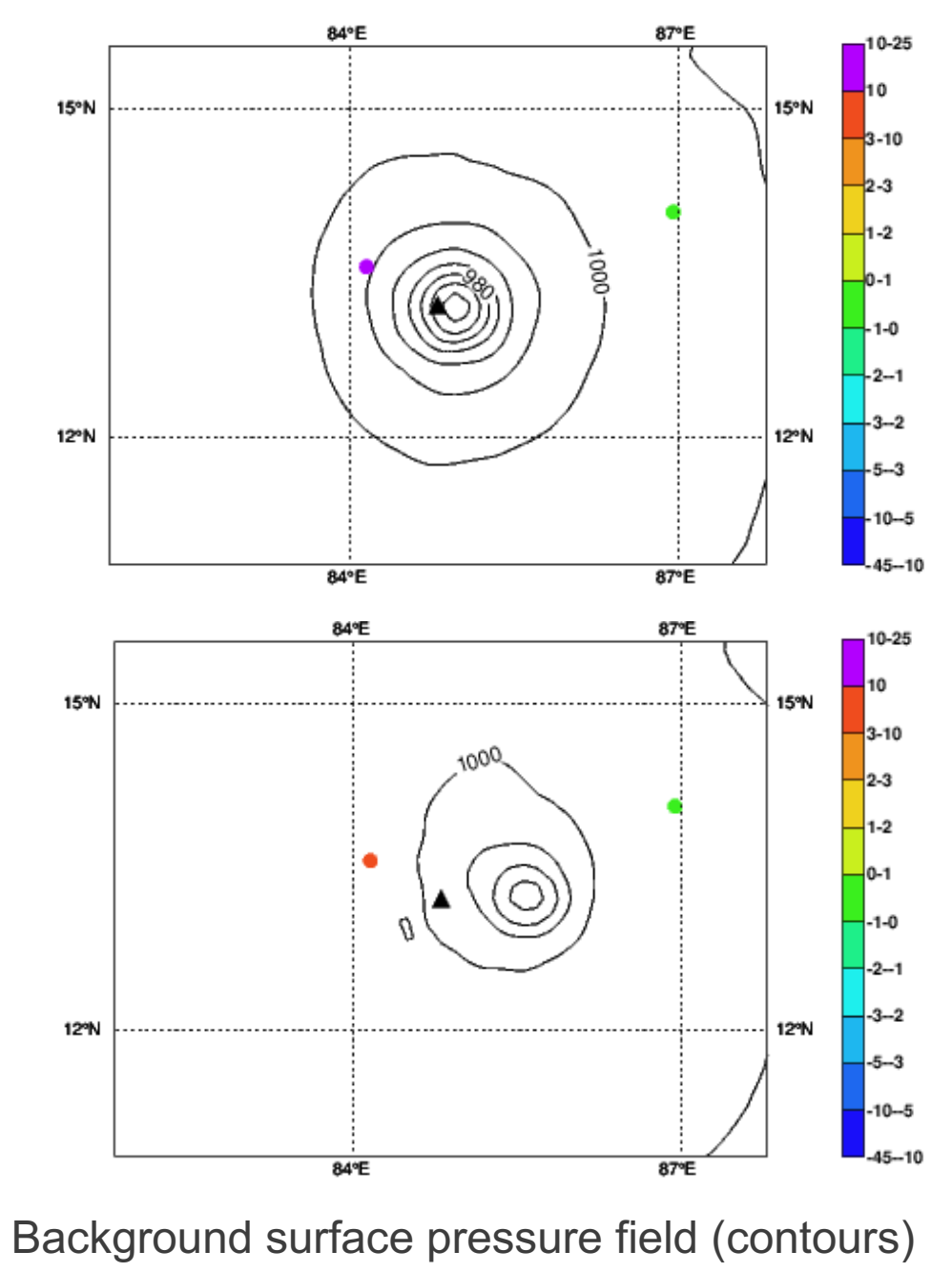


Figure 7: (top) Background surface pressure field (contours) and first guess departures of surface pressure (in hPa) plotted as filled circles. Analyzed field and analysis departures of surface pressure are shown in bottom panel. The statistics are on 30 April 2019.

Conclusion

This implementation of the machine learning data checking system aims to incorporate novel techniques in the detection and classification of observation anomalies. The new system tends to have fewer false alarms than the current operational framework, and it is able to detect all relevant anomalies. The classification of detected events is expected to improve with revised training based on improved labelling of past events.