DES HE VALP AI Convection Forecasts

Deliverable ID:	**D2.4**
Project Acronym:	**KAIROS**
Grant:	**101114701**
Call:	**HORIZON-SESAR-2022-DES-IR-01**
Topic:	**HORIZON-SESAR-2022-DES-IR-01-WA5-1**
Consortium Coordinator:	**Applied Innovative Methods**
Edition date:	**23 March 2024**
Edition:	**00.03**
Status:	**Official**
Classification:	**SEN**

Abstract

This document is a Validation Plan (VALP) for the AI Convection Forecast in the Industrial Research Fast Track Innovation & Uptake project, KAIROS. The KAIROS project aims at leveraging artificial intelligence to improve meteorological information for aviation stakeholders. This document briefly overviews the KAIROS project and proposes a plan for validating the project solution 1 AI convection forecast. This document will be updated periodically to capture the validation strategy throughout the execution of the KAIROS project. The current version of this document will only cover the validation strategy of the AI-based convection prediction forecast. The operational validation aspects of the solutions will be included in the next submission of the VALP. An incremental and final version of the VALP for Solution 1 will be submitted in October 2024.

EUROPEAN PARTNERSHIP

Co-funded by the European Union

Authoring & Approval

Author(s) of the document

Organisation name	Date
Universidad Carlos III de Madrid	March 6th, 2024
Applied Innovative Methods	March 10th, 2024

Reviewed by

Organisation name	Date
AI METHODS	March 15th, 2024
BIRA-IASB	March 15th, 2024
DSNA	March 15th, 2024
ENAIRE	March 15th, 2024
CRIDA	March 15th, 2024
EUROCONTROL	March 11th, 2024
FMI	March 15th, 2024
IGA	March 15th, 2024
METEOMATICS	March 15th, 2024
METSAFE	March 15th, 2024
UNISPHERE	March 15th, 2024
UC3M	March 15th, 2024

Approved for submission to the SESAR 3 JU by[1]

Organisation name	Date
AI METHODS	March 15th, 2024
BIRA-IASB	March 15th, 2024
DSNA	March 15th, 2024
ENAIRE	March 15th, 2024
CRIDA	March 15th, 2024
EUROCONTROL	March 15th, 2024
FMI	March 15th, 2024
IGA	March 15th, 2024

[1] Representatives of all the beneficiaries involved in the project

METEOMATICS	March 15th, 2024
METSAFE	March 15th, 2024
UNISPHERE	March 15th, 2024
UC3M	March 15th, 2024

Rejected by[2]

Organisation name	Date

Document History

Edition	Date	Status	Company Author	Justification
00.01	Jan 26th, 2024	Version 1	Javier García-Heras (UC3M)	Draft
00.02	March 10th, 2024	Update	Aniel Jardines (AI Methods)	Draft
00.03	March 23rd, 2024	Submission	Aniel Jardines (AI Methods)	Official

[2] Representatives of the beneficiaries involved in the project

KAIROS

UNLOCKING THE POTENTIAL OF AI-BASED WEATHER FORECASTS FOR OPERATIONAL BENEFITS

KAIROS

This document is part of a project that has received funding from the SESAR 3 Joint Undertaking under grant agreement No 101114701 under European Union's Horizon Europe research and innovation programme.

Table of Contents

List of Figures

List of Tables

1 Executive summary

KAIROS is a FTI&U project that promises to bring accurate and digital weather information to the aviation community. Solution 1 – AI Convection Forecast is focused on predicting convective weather, a significant cause of delays when the air traffic management system operates at maximum capacity. The technology is currently at TRL 3, and it is envisioned to reach TRL 7 by the end of the project. This document is the initial version of the validation plan that will be carried out to validate the technology and the benefits it brings to various stakeholders.

There are two main objectives in validating the AI Convection Forecast technology: **1) measure improvement of AI forecast at identifying areas of convective weather compared to conventional forecast**, and **2) assess the operational benefits of using AI Convection forecast to aviation end-users.** This initial version of the validation plan is focused on the first objective of quantifying the improvement in forecasting skills of AI convection forecast. A validation description of how the second objective related to assessing the operational benefits will be presented during the next iteration of the VALP deliverable in October 2024.

As the AI Convection Forecast technology matures from TRL3 to TRL7, validation activities are expected to be completed to achieve each maturity gate.

- **Maturity Gate TRL 4 – Historical Analysis – Spring 2024 -** The initial maturity gate entails a historical analysis of the model's performance on historical forecasts. We can compare the AI Convection model results with the convection "business as usual" forecasts based on historical data.
- **Maturity Gate TRL 6 – Real-Time Assessment – Summer 2024-**The following maturity gate will work with live forecast data of the AI Convection model. This activity will consist of real-time assessments of how the AI convection model results compare with conventional forecasts.
- **Maturity Gate TRL 7 – Operational Demonstration – Summer 2025**– The final maturity gate and one of the main goals of the KAIROS project is to use the AI convection forecast in an operation demonstration. Note that this validation activity will be part of the final version of the VALP, where the stakeholder's benefits are studied.

The validation activities are envisioned to assess the multiple types of forecasts produced by the AI Convection Forecast solution. Three different convection forecasts will be developed to meet the needs and requirements of various end users. Their spatiotemporal resolution and end user can categorise the three forecasts.

European Scale Convection – "Regional Forecast" - Pan-European forecast that will predict the convective situation over the entire ECAC region for the following 48 hours. The forecast has a spatial resolution of about 27 km and a temporal resolution of 1hr. Potential end user would be the Network Manager.

National Scale Convection – "Sub-Regional/National Forecast" – Sub-Regional forecast with a spatial resolution of ~13km, and temporal resolution will remain at 1hr. Potential end users would be the ANSP for national or cross-border use.

Local Scale Convection – "Local Forecast/Now-Cast" – High-resolution forecast/nowcast intended for local applications. The spatial resolution of the forecast will be 1km with a 20 min temporal resolution. Potential end users would be the Airport operators.

2 Introduction

2.1 Purpose of the document

This document provides the initial validation plan for KAIROS Solution 1 – AI Convection Forecast. It describes how stakeholder needs (currently being defined and formalised as a set of requirements within SPR-INTEROP/OSED, KAIROS D2.10) are intended to be validated.

Validation activities of solution 1 will be geared toward providing evidence that shows 1) the AI Convection Forecast provides better forecast skills over conventional forecasts in use today and 2) assesses the operational benefits to end users.

This initial version of the document is geared presenting the plan for validating the forecasting skill of the AI convection forecast compared with conventional convection forecast currently in operation. A follow-up version of this document (Oct 2024) will contain additional details on the validation plan for assessing the operation benefits to end users.

2.2 Intended readership

The intended readership of this validation plan is any stakeholder interested in learning how the technology will be validated and what evidence will be collected to demonstrate the merits of the solutions. The following is a preliminary list of stakeholders with possible interest in the KAIROS Solution 1 Validation Plan:

- **KAIROS Consortium** – Project members should read the VALP to ensure the plan is consistent with their planned activities within the KAIROS project.
- **SESAR Staff** – SESAR staff should read VALP to learn about validation activities and identify any potential connections with other ongoing projects.
- **Aviation End Users** – Potential end users of the technology should read the technology validation plan and help evaluate technical merit and identify any potential gaps from an operational perspective.
- **MET Providers** – Given the nature of the technology, current MET providers should read the plan validation plan and help evaluate technical merit and identify any potential gaps in assessing the skill of the AI Convection forecast.

2.3 Background

KAIROS continues the work set out in the SESAR Exploratory Research project, ISOBAR. ISOBAR consisted of integrating accurate and probabilistic convective weather forecasts within the Air Traffic Flow and Capacity Management (ATFCM) process. These weather forecasts were used as an input for demand and capacity prediction modules to provide early identification of imbalances between capacity and demand and anticipate adequate mitigation measures to ensure safety and maximise efficiency and capacity. ISOBAR defined an enhanced and highly automated ATFCM concept, exploiting AI to select mitigation measures at the local and network levels in a collaborative ATFCM operations paradigm.

The ISOBAR project developed the “MetEngine”, an AI-based algorithm for predicting convective areas impacting air traffic flow management (ATFM) operations. KAIROS Solution 1 is focused on further developing the “MetEngine”, by increasing its technology readiness level and making the transition from research to uptake by industry.

2.4 Structure of the document

The document is structured as follows:

- Chapter 2, “Introduction”, describes the purpose of the document, the intended readership and the background and gives an explanation of the abbreviations and acronyms used throughout the document.
- Chapter 3, “Context of the Validation”, deals with the context of the validation and provides a summary of the solutions implemented for validation. A list of stakeholders with need and involvement is provided.
- Chapter 4, “Validation Approach”, focuses on the validation approach, the stakeholder's expectations, and validation objectives in the main performance areas identified for the project.
- Chapter 5, “Validation Activities”, details the assumptions and provides a description of the reference scenario and of the validation exercises. Each exercise is described as well as the planning and the validation platform.
- Chapter 6, “References”, lists all the applicable and reference documents.

2.5 Glossary of terms

The list of terms will be updated in future versions of the VALP document.

Term	Definition	Source of the definition
		3

Table 1: glossary of terms

2.6 List of acronyms

Acronym	Description
AI	Artificial Intelligence
ANSP	Air Navigation Service Providers
ATM	Air Traffic Management
ATFM	Air Traffic Flow Management
CDE	Communication, Dissemination, and Exploitation
CNN	Convolutional Neural Network
DOI	Digital Object Identifier
ER	Exploratory Research
EU	European Union
FAIR	Findable, Accessible, Interoperable, Reusable
FMP	Flight and Meteorological Planning
FMP	Flow Manager Position
FN	False Negatives
FP	False Positives
FPR	False Positive Rate
HRB	Horizon Results Booster
HRP	Horizon Results Platform
IPR	Intellectual Property Rights
JU	Joint Undertaking
KoM	Kick-off Meeting
KPIs	Key Performance Indicators
LSTM	Long Short-Term Memory
NWP	Numerical Weather Prediction
PPI	Plan Position Indicator
RDT	Rapid Development Thunderstorm
ROC	Receiver Operating Characteristic
SESAR	Single European Sky ATM Research
SMEs	Small and Medium-sized Enterprises
TN	True Negatives
TP	True Positives
TPR	True Positive Rate

Table 2: List of acronyms

3 Context of the validation

3.1 Validation plan context

The KAIROS Solution 1 Validation Plan aims to describe the validation context, exercises, and activities that will take place to validate the artificial intelligence-based convection forecast concept. The current solution will rely on three different AI models based on the spatiotemporal resolution provided, ranging from 24 km over 1 hour (European Scale) to 13 km and 1 hour (National Scale), culminating in a model capable of predicting with a resolution as fine as 1 km and 20 minutes (Local Scale). Three validation exercises are envisioned to focus on validating the three AI Convection Models, progressing from low to high spatiotemporal resolutions. The Regional Scale model will undergo validation in EXE01, the Sub-Regional Scale model in EXE02, and the Local Scale model in EXE03. The validation exercises will be carried out in multiple phases and in accordance with the two main validation objectives: 1) the AI Convection Forecast provides better forecast skills over conventional forecasts in use today and 2) assesses the operational benefits to end users.

Only objective 1 will be addressed in this current version of the VALP, an incremental and final version of the VALP is expected for October 2024, and will cover the validation of the expected operational benefits of KAIROS solution 1.

For ***Validation Exercise #1-Regional Scale model***, please refer to Section 5.1. This exercise will incorporate data from NWP forecasts and observations covering the region of Europe (lat[20,70] and long[-20, 40]). Initially, a historical analysis will be conducted to demonstrate the model's performance metrics utilizing data from the summer of 2023. Following this historical analysis, the AI forecast skill will be assessed continuously by comparing the model prediction with the observation data from satellites, lightning, and radar. After the historical assessment, a real-time evaluation of the AI model will be done.

For ***Validation Exercise #2-Sub-Regional Scale model***, refer to Section 5.2. Similar to EXE#1, EXE#2 will perform an initial historical analysis utilizing forecast and observation data from 2023 to quantify the models' forecasting skill. This exercise will focus on the sub-region of Western Europe, focusing on Spain and France. This historical analysis will be performed in Spring 2024. Following this historical assessment, the model performance skill will be evaluated continuously in real-time starting in summer 2024 and throughout the duration of the KAIROS project.

Finally, the ***Validation Exercise #3-Local Scale model***, refer to Section 5.3. This exercise will validate the AI convection "now-casting" model at a local scale and is specifically tailored for airport operations. Consequently, it requires access to local weather data, high-resolution EPS NWP forecasts, and thunderstorm observations using radar and lightning. As in the other exercises, the model data will be validated using historical data from 2023 and followed by real-time assessment of forecasting skill using weather observations.

3.2 KAIROS solution 1 "AI Convection Forecast": a summary

KAIROS Solution 1 will focus on raising the TRL of the "MetEngine", an AI-based algorithm developed within the SESAR ER project ISOBAR to predict convective areas impacting air traffic flow management (ATFM) operations. Within the ISOBAR project, validation exercises were conducted using historical data, resulting in positive feedback from operators. The current maturity level of this solution is TRL 3;

the expectation within KAIROS is to demonstrate the technology within an operational environment to reach TRL 7.

SESAR solution ID	SESAR solution title	SESAR solution description	Enabler ref. (from SESAR architecture)	Enabler coverage
KAIROS Solution 1	AI-based Convection Prediction	Integration of AI-based convection prediction models within air traffic flow management (ATFM) operational tools and platforms. This solution will provide convection prediction customizable to meet the requirements for the pre-tactical and tactical processes of ATFM.	Enabler 1 "MetEngine"	Required & Use: The enabler will be used as a starting point in an AI tool's development process to provide enhanced forecast convection.

Table 3: SESAR solution under validation

3.3 KAIROS solution 1 "AI Convection Forecast": key R&I needs

Weather is inherently a capacity and demand issue as severe events reduce the airspace available to aircraft. KAIROS will contribute to the strategic research and innovation agenda of SESAR 3 JU. The project mainly addresses the needs of the R&I flagship's *capacity on demand and dynamic airspace* by providing a more accurate weather prediction to support pre-tactical and tactical ATFM operations. KAIROS will also touch on elements of Artificial intelligence for aviation, as it aims to deploy an AI-model within an operational environment. Specific *R&I needs* that KAIROS Solution 1 will cover include:

1. **On-demand ATS**: KAIROS will improve the quality and format of weather and capacity forecasts, increasing the network stakeholders' confidence in planning information.
2. **ATM continuity of service despite disruption**: KAIROS will provide a smart digital solution capable of predicting adverse weather situations impacting airspace operations. Weather forecasts produced within the KAIROS Solution 1 will be provided in a digital format to enable integration with a suite of existing and novel tools to facilitate decision-making processes such as DCB and trajectory planning by multiple stakeholders during weather-related disruption scenarios.
3. **Future data services and applications for airport and network**: The main objective of the KAIROS project is to mature the TRL level of AI-based weather forecasting to TRL 7 (operational demonstration). The intention is to create a digital weather service available to various aviation stakeholders at the airport and network levels.
4. **Trustworthy AI-powered ATM environment**: The KAIROS project will perform analysis to provide explainable additional validation activities, including comparison with conventional baselines and online learning assurance, to provide a transparent, robust, and stable solution under all conditions.
5. **Human–AI collaboration**: The KAIROS project will address how to best display AI-based weather information and suggest actions with aviation actors. This work will contribute to a

better understanding of how humans and AI applications can collaborate in the early prediction/detection of weather-related risks.

The research questions extracted from the previous detailed R&I needs can be summarised in the following table:

Research Question #01	How can the KAIROS AI convection forecast enhance the network stakeholders' confidence in planning information?	R&I#01
Research Question #02	How can KAIROS AI convection forecast improve the quality of the existing weather forecast?	R&I#01
Research Question #03	Which is the KAIROS AI convection forecast format that augments the format used in the existing weather forecast?	R&I#01; R&I#02
Research Question #04	How can the KAIROS AI convection forecast solution integrate with existing and novel tools to ensure ATM continuity of service?	R&I#02; R&I#03
Research Question #05	What specific data services and applications can be developed for airports and network stakeholders through the KAIROS AI convection forecast solution?	R&I#03
Research Question #06	How can the KAIROS AI convection forecast solution be effectively tailored and made accessible to various aviation stakeholders operating at both airport and network levels?	R&I#03
Research Question #07	How can the KAIROS AI convection forecast solution ensure transparency, robustness, and stability under various operational conditions?	R&I#04
Research Question #08	How can the KAIROS AI convection forecast solution be implemented to enhance explainability and trustworthiness?	R&I#04
Research Question #09	What are the efficient methods for displaying the KAIROS AI convection forecast solution and suggesting actions to aviation actors?	R&I#05
Research Question #10	How can the collaboration between humans and AI applications be optimised for early prediction and detection of weather-related risks in aviation operations?	R&I#05

3.4 KAIROS solution 1 "AI Convection Forecast" Estimated Performance Contributions (EPC)

From ATM master plan 2020 edition: Interoperable digital AIM and MET services are an essential prerequisite for TBO and, therefore, need to be deployed before TBO. KAIROS solution 1 aims to provide an enhanced convection forecast based on AI that will contribute to impacting the following

KPAs. The Estimated Performance Contributions of KAIROS Solution 1 are summarized in the table below. The numbers under Relative Impact are indicative of relative impact with respect to all the other SESAR solutions contributing to that specific KPA (1 for low, 2 for medium, 3 for high).

Table 4: KAIROS Solution 1 – Estimated Performance Contributions

Abbreviation	Description	Relative Impact
SAF	Safety	Indirect Safety Impact
FEFF1	Actual average fuel burn per flight	2 - Medium
CAP2	En-route throughput, in challenging airspace, per unit time	2 - Medium
PUN1	Average departure delay per flight	2 - Medium
HP	Human Performance	Yes
DIGI	Digitalisation	Yes

3.5 Initial and exit maturity levels

The KAIROS solution 1 “AI convection forecast” starts at TRL level 3 and is intended to reach maturity level TRL7 by the end of the project. Each of the three models produced will be validated in a dedicated exercise.

KAIROS solution 1	SESAR solution title	Initial maturity level	Exit maturity level	Reused validation material from past R&I Initiatives
KAIROS SOLUTION 1	AI convection forecast	TRL 3	TRL 7	None

Table 5: maturity levels table

4 KAIROS solution 1 “AI Convection Forecast” validation plan

4.1 Validation approach

This section describes the validation approach of the KAIROS project. It provides the list of the different validation exercises and the links to the research questions in Section 3.3 and KPAs in Section 3.4. The approach of validating KAIROS Solution 1 aligns with the project’s overall objectives: **1**) measure improvement of AI forecast at identifying areas of convective weather compared to conventional forecast, and 2) assess the operational benefits of using AI Convection forecast to aviation end-users.

The initial validation phases will be geared towards validating the forecasting skill of the AI models (Objective 1), while later phases will aim at validating the stakeholders’ performance benefits of the solution. The detailed information of further phases in the exercises will be updated in future versions of the VALP document, which are expected in October 2024.

ID	Title	Exit TRL	RQ	KPA	KPI
EXE01.1	Regional Forecast historical analysis based on AI performance metrics	TRL 4	RQ#02; RQ#07		ROC, AUC,
EXE01.2	Regional Forecast real-time analysis based on AI performance metrics	TRL 6			
EXE02.1	Sub-Regional/National Forecast historical analysis based on AI performance metrics	TRL 4	RQ#02; RQ#07		ROC, AUC,
EXE02.2	Sub-Regional/National Forecast real-time analysis based on AI performance metrics	TRL 6			
EXE03.1	Local Forecast historical analysis based on AI	TRL 4	RQ#02; RQ#07		ROC, AUC,

	performance metrics				
EXE03.2	Local Forecast real-time analysis based on AI performance metrics	TRL 6			

4.2 Stakeholders' expectations and involvement

Because the initial VALP is focused on validating the models without incorporating the benefits analysis for stakeholders that is expected for the final version of the VALP, which is expected for October 2024, no stakeholders are involved in the validation exercises.

4.3 Validation objectives

The following table only lists the validation objectives for the models developed for the KAIROS solution 1; the operational objectives will be added in the final VALP, where the stakeholder benefits analysis will be performed.

Objective ID	Objective title	Objective description	Success Criteria	Research Questions
O1	Produce AI-based MET forecast	Apply artificial intelligence algorithms on available forecast and observation weather data to improve the prediction of several weather phenomena impacting aviation (convective weather).	AI algorithms show improvement (accuracy and lead time) with respect to weather information available today. Deployment of AI-based MET models on live data to produce operational forecasts	From RQ#01; to RQ#10
O2	Accuracy of prediction	Provide evidence that the prediction provided by the AI Convection Forecast tool produces accurate predictions of convective storms	Predicted 2D convective area with severity 1 at lead time -48h differs from observed area in less than 30%. Idem for other combinations of severity and lead times. The prediction of convective weather of severity 1 is accurate at least 90% of the times.	RQ#02

O3	Accuracy improvement	Provide evidence that the prediction provided by the AI Convection Forecast tool produces predictions of convective storms that are more accurate than existing tools available today	Similar to the above SC but in comparison with today's tools/ predictions	RQ#01; RQ#02; RQ#07; RQ#08
O4	Forecast format	Check that the AI Convection Forecast can produce in real time predictions of convective storms adapted to the format required by the end-users	Output data of the tool is provided in GeoJSON, WFS or WCS formats including information for each cell in the airspace of 2,5 Km2 and each FL between 100 and 460 about probability of convective event, severity and presence of lightning. The data must also indicate the time of forecast and time when the convective phenomenon is expected to occur.	RQ#03;
O5	Number of predictions	Demonstrate that the AI Convection Forecast can produce forecasts at various leadtimes	The tool provides forecast at leadtime -48h The tool provides forecast at leadtime -36h Etc.	RQ#01;
O6	Quality of predictions	Check that minimum data quality process is applied to the AI Convection Forecast input data and development process	Input data is documented, including sources, formats and resolution; Data annotation process is automated and documented; Traceability of input data from source to pipeline is documented; Mechanisms are in place to prevent input data corruption during storage or processing; The level of independence between training, validation and test datasets is documented.	RQ#02;
O7	Real-time performanc	Validate that the AI Convection Forecast is	Here we need to stablish the nominal required (by end-users) level of	RQ#02; RQ#04;

	e monitoring	able to perform real-time performance monitoring and alert in case the predictions produced have a level of confidence below the nominal standards	confidence and the necessary indicators that the model have to monitor in real-time to warn end-users about possible under-performance of the model (which can be due to internal mechanisms or to lack of quality of input data). The nominal performance does not necessarily match the SC proposed for the VAL Obj related to "Accuracy of prediction", since those SC represent the "ideal" performance of the model, whereas the real-time performance monitoring can alert only in cases where performance is not only non-ideal, but also poor.	RQ#07; RQ#08;
O8	Post-ops analysis	Validate that the AI Convection Forecast is able to perform post-operational performance analysis	Storage module is in place; Post-operational calculations over the data stored are automated; Reports are produced when post-operational analysis detect under-performance (in line with VALObj "Real-time performance monitoring"), including information about time of the prediction, leadtime, input data used, output prediction, actual observation, possible malfunctions impacting model performance, etc.	RQ#01; RQ#02; RQ#07; RQ#08

4.4 Validation assumptions

Assumption ID	Assumption title	Assumption description	Justification	Impact Assessment
A#01	Storm observation	While we operate under the assumption that the data supplied by the RDT product is entirely accurate, there is a prevailing belief that the RDT product tends	Although errors are anticipated within the RDT product, it represents the most comprehensive and reliable information currently available	KAIROS models will be biased towards the RDT observations

		to overestimate the severity of storms.	regarding convective weather conditions.	

Table 6: validation assumptions overview

4.5 Validation exercises list

KAIROS is composed of 3 validation exercises, as listed below. They are further detailed in Chapter 5.

ID	EXE01
Title	European Scale Convection – Regional forecast
Description	This exercise will focus on predicting the convective situation over the entire ECAC region for the following 48 hours. The potential end user would be the Network Manager.
Expected Achievements	Convective weather prediction with a spatial resolution of about 27 km and a temporal resolution of 1hr.
Use Cases	Phase#01. Historical Analysis. The initial maturity gate entails a historical analysis of the model's performance on historical forecasts. We can compare the AI Convection model results with the convection “business as usual” forecasts based on historical data. Phase #02. Real-Time Assessment. The following maturity gate will work with live forecast data of the AI Convection model. This activity will consist of real-time assessments of how the AI convection model results compare with conventional forecasts. Real time assessment of AI forecasts will be performed in an automatic and continuous manner throughout the duration of the KAIROS project.
Validation Technique	Phase#01. Literature Study and Judgemental Techniques Phase #02. Real-time simulation
KAP/TA Addressed	
Start Date	Phase#01. 01/04/2024 Phase#02. 15/05/2024
End Date	Phase#01. 30/11/2024 Phase#02. 31/05/2026
Validation Coordinator	AI Methods

Validation Platform	Phase#01 and #02. Validation Platform Models will be evaluated using python programming language.
Validation Location	Phase#01 and #02. AI Methods, Leganés, Madrid (Spain).
Dependencies	None

ID	EXE02
Title	National Scale Convection – Sub-Regional forecast
Description	This exercise will be focused on predicting the convective situation over a sub-Regional area in Europe. Potential end users would be the ANSP for national or cross-border use.
Expected Achievements	Convective weather prediction with a spatial resolution of about 13km and a temporal resolution of 1hr.
Use Cases	Phase#01. Historical Analysis. The initial maturity gate entails a historical analysis of the model's performance on historical forecasts. We can compare the AI Convection model results with the convection "business as usual" forecasts based on historical data. Phase #02. Real-Time Assessment. The following maturity gate will work with live forecast data of the AI Convection model. This activity will consist of real-time assessments of how the AI convection model results compare with conventional forecasts. Real time assessment of AI forecasts will be performed in an automatic and continuous manner throughout the duration of the KAIROS project.
Validation Technique	Phase#01. Literature Study and Judgemental Techniques Phase #02. Real-time simulation
KAP/TA Addressed	
Start Date	Phase#01. 01/04/2024 Phase#02. 15/05/2024
End Date	Phase#01. 30/11/2024 Phase#02. 31/05/2026
Validation Coordinator	AI Methods

Validation Platform	Phase#01 and #02. Validation Platform Models will be evaluated using python programming language.
Validation Location	Phase#01 and #02. AI Methods, Leganés, Madrid (Spain).
Dependencies	None

ID	EXE03
Title	Local Scale Convection – Local Forecast/nowcast
Description	High-resolution forecast/nowcast intended for local applications. Potential end users would be the Airport operators.
Expected Achievements	Convective weather prediction with a spatial resolution of about 1km and a temporal resolution of 20 min.
Use Cases	Phase#01. Historical Analysis. The initial maturity gate entails a historical analysis of the model's performance on historical forecasts. We can compare the AI Convection model results with the convection “business as usual” forecasts based on historical data. Phase #02. Real-Time Assessment. The following maturity gate will work with live forecast data of the AI Convection model. This activity will consist of real-time assessments of how the AI convection model results compare with conventional forecasts. Real time assessment of AI forecasts will be performed in an automatic and continuous manner throughout the duration of the KAIROS project.
Validation Technique	Phase#01. Literature Study and Judgemental Techniques Phase #02. Real-time simulation
KAP/TA Addressed	
Start Date	Phase#01. 01/06/2024 Phase#02. 15/07/2024
End Date	Phase#01. 31/12/2024 Phase#02. 31/05/2026
Validation Coordinator	AI Methods
Validation Platform	Phase#01 and #02. Validation Platform Models will be evaluated using python programming language.

Validation Location	Phase#01 and #02. AI Methods, Leganés, Madrid (Spain).
Dependencies	None

4.6 Validation exercises planning

The table below provides a general timeline for the Phase 1 activities within each exercise.

ID	Q1 2024			Q2 2024			Q3 2024			Q4 2024			Q1 2025		
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar
EXE01															
EXE02															
EXE03															

4.7 Deviations with respect to the SESAR 3 JU project handbook

No deviations have been identified from SESAR 3 Project handbook.

Co-funded by the European Union

5 KAIROS Validation exercises

5.1 Validation Exercise #01 plan. European Scale Convection – Regional forecast

5.1.1 Validation exercise description and scope

This exercise is designed to validate the effectiveness of the KAIROS AI convection forecast at the network level, specifically focusing on its impact on storm forecasts on the Network Manager (NM) planning decisions. The AI convection forecast model utilises a sophisticated architecture: the Convolutional Neural Network combined with Long Short-Term Memory (CNN-LSTM). This model is trained using the following observational data:

- Rapid-Development Thunderstorm (RDT),
- Lightning detection data
- Satellite-based Cloud Top Height Observations

RDTs provide essential information, including storm contours, overshot locations, severity, cloud top details, velocity, and direction. Integrating these diverse data points enhances the model's ability to deliver accurate and comprehensive convection forecasts.

The model's geographical scope extends to the European region, with a forecast temporal range spanning 48 hours.

This exercise aims to demonstrate that the KAIROS AI convection Regional forecast improves current products in refining storm predictions and supporting strategic decision-making for Network Manager planning. It will be developed in 2 sequential phases:

1. **Phase#01: Model Performance Demonstration**. The initial phase focuses on demonstrating the model's performance in storm predictions. Specific statistical metrics will be utilised to assess the accuracy and reliability of storm predictions. This phase will validate the performance of the AI models on a historical dataset.

2. **Phase#02:** Real-Time **Analysis.** The second phase entails showcasing the forecast model's tangible benefits for real-time applications. Through comprehensive analysis, the exercise aims to highlight the value of the KAIROS solution for stakeholders in a future real application.

5.1.2 Stakeholders' expectations and benefit mechanisms addressed by the exercise.

Stakeholder	Involvement	Why it matters to the stakeholder
NM	No active involvement in this initial VALP.	Increased thunderstorm activity throughout Europe has been a significant cause of ATFM delays in recent years. Bad weather coupled with an increasing demand on the system causes significant disruption to ATFM operations. Weather is typically handled

		tactically by local FMPs; however, decisions made at the local level can lack the network-wide perspective. Better weather information earlier in the ATFM planning process would enable the NM to take strategic actions at the network level.

Table 7: stakeholders' expectations

5.1.3 Validation objectives

The validation objectives for the current validation plan are focused on validating the skill of the convection forecasting algorithm.

Objective	Explanation
Objective #01: Produce AI-based MET forecast	Apply artificial intelligence algorithms on available forecast and observation weather data to improve the prediction of several weather phenomena impacting aviation (convective weather).
Objective #02: Accuracy of prediction	Provide evidence that the prediction provided by the AI Convection Forecast tool produces accurate predictions of convective storms
Objective #03: Accuracy improvement	Provide evidence that the prediction provided by the AI Convection Forecast tool produces predictions of convective storms that are more accurate than existing tools available today
Objective #04: Forecast format	Check that the AI Convection Forecast is able to produce in real time predictions of convective storms adapted to the format required by the end-users
Objective #05: Number of predictions	Demonstrate that the AI Convection Forecast can produce forecasts at various leadtimes
Objective #06: Quality of predictions	Check that minimum data quality process is applied to the AI Convection Forecast input data and development process
Objective #07: Real-time performance monitoring	Validate that the AI Convection Forecast is able to perform real-time performance monitoring and alert in case the predictions produced have a level of confidence below the nominal standards
Objective #08: Post-ops analysis	Validate that the AI Convection Forecast is able to perform post-operational performance analysis

5.1.4 Validation scenarios

In Phase#01, the validation process will use historical data from 2023 to formulate scenarios for historical analysis. Note that 2023 data was not included in the model development datasets, so the model has never viewed the scenarios used in the validation process. Each scenario will undergo rigorous validation across multiple forecast releases within the KAIROS solution, ensuring a comprehensive assessment of its predictive capabilities. Validation scores will be meticulously derived for each storm characteristic, allowing for detailed evaluation. In Phase#02, as a real-time analysis, the weather data will be from the day of the exercise execution. The area of interest for KAIROS solution 1 Regional model will cover the region of Europe (latitude $\in$[20,70] and longitudes $\in$ [-20, 40], see figure below).

Figure 1: KAIROS Solution 1 regional model geographic domain

5.1.4.1 Reference scenario(s)

The main reference for comparing the performance of the model will be the observation data. Observational data will allow for characterising the occurrence, severity and altitude of storms. Data to be used includes RDT, lightning strikes, and Cloud Top height. Additionally, the solution will utilize current convection forecasts, such as the Cross-Border Initiative, as a reference for comparing model performance.

5.1.4.2 Solution scenario(s)

Solution scenarios are obtained by applying the KAIROS solution, obtaining AI-enhanced occurrence, severity and altitude of storms and occurrence of lightning.

We should then define indicators to assess how AI-enhanced weather forecasting (occurrence, severity, altitude, lighting) can predict the reference scenario. We will use traditional ML metrics; for

more information, see Appendix A, e.g., Receiver Operating Characteristic (ROC) curve, Precision-Recall (PR) curve, and Confusion Matrix.

5.1.5 Exercise validation assumptions

No additional assumptions apply to this exercise; see Section 4.4.

5.1.6 Limitations and impact on the level of significance

An inherent limitation in our validation strategy resides in exclusively utilising RDT data as the primary ground truth for training the KAIROS solution 1 model. While RDT data provides invaluable insights into thunderstorms, it may not have all nuances of storm behaviours germane to stakeholder considerations. Storm duration, size, intensity fluctuations, and localised phenomena can profoundly affect stakeholders' perceptions of storm severity and decision-making paradigms. This constraint may generate disparities between the model's predictions and stakeholders' specific expectations.

We propose a dedicated tool to mitigate this constraint and enhance the fidelity of the KAIROS solution 1 model to stakeholders' requirements. This tool will empower stakeholders to set threshold values more accurately and encapsulate their criteria for characterising storm magnitude or severity. By enlisting stakeholder input in the selection process, we aim to refine the utility of the model's prediction in operational settings. This collaborative paradigm will engender storm predictions from the KAIROS solution 1 model that is more attuned to stakeholders' requirements and expectations.

5.1.7 Validation exercise platform/tool and validation technique

5.1.7.1 Validation exercise platform / tool characteristics

KAIROS solution 1 will be developed in Python, using the Keras library to implement the artificial recurrent neural network architecture. Every part of the AI model will be created using in-house computers. A dedicated platform will also be developed to show and study the model's performance metrics. The description of the platform can be found in Appendix B of the document.

5.1.7.2 Validation exercise Technique

The KAIROS Convection model will be validated against convection observation data, including lightning, satellite, and radar. The validation process employs advanced machine learning metrics; see Appendix A for more detailed information, namely here the most representative of the ROC curve and confusion matrix. The ROC curve, short for Receiver Operating Characteristic Curve, offers a visual depiction of the True Positive Rate (TPR), which signifies the probability of detection, juxtaposed with the False Positive Rate (FPR), indicative of false alarms across diverse threshold settings. Meanwhile, the confusion and error matrices provide a comprehensive visualisation of algorithmic performance in supervised learning. It delineates the percentages of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), offering critical insights into model accuracy and efficacy.

5.1.8 Data collection and analysis

5.1.8.1 Data and data collection methods

The data utilised in this validation endeavour comprise the following components: RDT, lightning strikes, Cloud Top height and NWP forecasts supplementing the meteorological datasets.

To facilitate the validation process, distinct sets for training, validation, and testing are arranged in the combination of NWP forecasts and RDT satellite images with information from 2019 to 2022. This integration involves aligning the NWP grid with the higher-resolution satellite images and delineating the grid points within the RDT storm polygons. Interoperability between datasets is enhanced by standardising the data to a spatial resolution of .25 x .25 degrees.

Given the variance in temporal resolution (1 hour for NWP predictions compared to 15 minutes for RDT observations), an approach is devised to address this incongruity. A grid point is classified as convective if a storm observation is recorded during any of the four observation instances within the hour. This method enables constructing a binary training target function indicative of storm cell occurrence at a grid location within the hour.

Considering our focus on a 48-hour time horizon and the release of forecasts every 6 hours, disparate range forecasts valid for the same time frame are utilised for training, validation, and testing the model. This strategic adaptation accommodates the temporal dynamics inherent in the forecast data.

For the data collection regarding the validation exercise data (data from 2023), we will approach it similarly.

5.1.8.2 Analysis methods

Statistical analysis will be employed to evaluate the efficacy of the KAIROS solution 1 models, focusing specifically on the ML indicators, including the ROC curve and Confusion Matrix; see Appendix A for more information and indicators. Findings will be compared with observational data from convective storm and lightning observations sourced from the RDT product and lightning detection records.

5.1.9 Exercise planning and management

5.1.9.1 Activities

- Activity 1 - preparation of scenarios: Collect all weather data sources for the selected dates.
- Activity 2 – compute reference datasets: Prepare datasets with the target variables from the weather data sources.
- Activity 3 - compute model datasets: Extract dataset values for the selected dates.
- Activity 4 – data comparison and analysis of the results.
- Activity 5 – prepare the validation report.

5.1.9.2 Roles and responsibilities in the exercise

AI Methods will lead and develop the validation activities, Meteomatics will provide the NWP Forecasts products, RDTs are provided by MetSafe, lightning by MetSafe and Meteomatics and Cloud Top Height by the European Space Agency.

5.1.9.3 Time planning

ID	Q1 2024	Q2 2024	Q3 2024	Q4 2024

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Activity 1												
Activity 2												
Activity 3												
Activity 4												
Activity 5												

Table 8: Detailed planning for EXE01 Phase 1

5.1.9.4 Identified risks and mitigation actions

Risks	Impact (1-low, 2-medium, 3-high)	Likelihood (1-low, 2-medium, 3-high)	Criticality (calculated based on likelihood and impact)	Mitigation actions
Risk #01: AI models do not provide results: AI models may not be able to capture the complex dynamics of certain weather phenomena.	High	Low	High	Previous research has shown that AI is able to generalize the complex behaviour of extreme weather events. AI algorithm development will look to past research to design AI models and algorithms
Risk #02: Data availability: Not having the necessary data will hinder attempts to develop AI models. It is important to have a large volume of relevant data in a timely manner to produce the models according to the project timeline.	High	Low	High	Required data sources have been identified. Most of the data will be available via the project partners (Meteomatics, MetSafe). The project will also need to purchase certain data sets, the sources of these additional data sets have

				been identified and initial cost estimates have been obtained.

Table 9: exercise #01 risks and mitigation actions

5.2 Validation Exercise #02 plan. National Scale Convection – Sub-Regional forecast

5.2.1 Validation exercise description and scope

This exercise aims to validate the effectiveness of the KAIROS AI convection forecast, specifically focusing on its impact on storm predictions at the Sub-Regional level and its support for strategic decision-making by Air Navigation Service Providers (ANSPs). The AI convection forecast model utilises a sophisticated architecture, combining Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM). CNN and LSTM are artificial neural networks commonly used in pattern recognition and sequence prediction tasks.

As in the previous exercise, this model is trained using Rapid-Development Thunderstorms (RDT), lightning strikes, and Cloud Top height. Radar Plan Position Indicator Product (PPI) is also used to improve storm observation.

The forecast temporal range spans 48 hours, and its geographical scope extends to the FIR regions. The difference with the previous exercise is that the forecast employed as inputs for this model will include high-resolution local NWP. Through this exercise, our objective is to demonstrate the efficacy of the KAIROS AI convection forecast in refining storm predictions and supporting strategic decision-making for ANSP resources, including staff and sector configuration. This exercise will also be executed in 2 consecutive phases: the first phase will assess the model's accuracy with historical data, followed by activities to evaluate its behaviour on real-time applications. Please note that this initial version of the document will not incorporate the validation activities to analyse the effectiveness of the KAIROS solution in regular operations, focusing on its accuracy, reliability, and efficiency; those activities will be included in the final version of the VALP.

5.2.2 Stakeholders' expectations and benefit mechanisms addressed by the exercise.

Stakeholder	Involvement	Why it matters to the stakeholder
ANSP	No active involvement in this initial VALP.	Unforeseen weather phenomena emerge as formidable obstacles for air traffic control systems. Airborne traffic amidst extreme weather events demands meticulous coordination, as flights are forced to deviate from planned trajectories, navigate through holding patterns, and optimize fuel management strategies while seeking refuge at alternate airfields. Ground operations are similarly impacted, with flights susceptible to significant delays and cancellations, that may cascade throughout the network.

5.2.3 Validation objectives

Objective	Explanation
Objective #01: Produce AI-based MET forecast	Apply artificial intelligence algorithms on available forecast and observation weather data to improve the prediction of several weather phenomena impacting aviation (convective weather).
Objective #02: Accuracy of prediction	Provide evidence that the prediction provided by the AI Convection Forecast tool produces accurate predictions of convective storms
Objective #03: Accuracy improvement	Provide evidence that the prediction provided by the AI Convection Forecast tool produces predictions of convective storms that are more accurate than existing tools available today
Objective #04: Forecast format	Check that the AI Convection Forecast can produce in real time predictions of convective storms adapted to the format required by the end-users
Objective #05: Number of predictions	Demonstrate that the AI Convection Forecast can produce forecasts at various leadtimes
Objective #06: Quality of predictions	Check that minimum data quality process is applied to the AI Convection Forecast input data and development process
Objective #07: Real-time performance monitoring	Validate that the AI Convection Forecast is able to perform real-time performance monitoring and alert in case the predictions produced have a level of confidence below the nominal standards
Objective #08: Post-ops analysis	Validate that the AI Convection Forecast is able to perform post-operational performance analysis

5.2.4 Validation scenarios

In Phase#01, the validation process will employ historical data from 2023 to formulate scenarios for analysis. Each day within these scenarios, rigorous validation will be performed across multiple forecast releases within the KAIROS solution, ensuring a comprehensive assessment of its predictive capabilities. Validation scores will be meticulously derived for each storm characteristic, enabling detailed evaluation. In Phase#02, as a real-time analysis, the weather data will be from the day of the exercise execution. The area of interest for the KAIROS solution 1 Sub-Regional model will cover the region of Western Europe.

Figure 2: KAIROS Solution 1 sub regional model geographic domain

5.2.4.1 Reference scenario(s)

We should characterise the occurrence, severity and altitude of storms and the occurrence of lightning.

5.2.4.2 Solution scenario(s)

Using artificial intelligence techniques, solution scenarios are generated by applying the KAIROS solution, which enhances the prediction of storm occurrence, severity, altitude, and lightning. Following this, it is necessary to establish indicators to evaluate the predictive capability of AI-enhanced weather forecasting against the reference scenario. Traditional machine learning metrics, including the Receiver Operating Characteristic (ROC) curve, Precision-Recall (PR) curve, and Confusion Matrix, will be employed for this purpose. Further details can be found in Appendix A.

5.2.5 Exercise validation assumptions

No additional assumptions apply to this exercise; see Section 4.4.

5.2.6 Limitations and impact on the level of significance

As in the previous exercise, when RDT is used as ground truth, the model results tend to overestimate storm sizes, which could imply inefficiency for ANSP activities such as planning ATC sector

configuration or ATC tactical interventions. To better characterise storms, KAIROS models will be incorporated into the training phase storm observation from radar data.

5.2.7 Validation exercise platform/tool and validation technique

5.2.7.1 Validation exercise platform / tool characteristics

KAIROS solution 1 will be developed in Python, using the Keras library to implement the artificial recurrent neural network architecture. Every part of the AI model will be created using in-house computers. A dedicated platform has also been devised to present the model's performance metrics. Appendix B illustrates a preliminary version of this tool.

5.2.7.2 Validation exercise Technique

The validation process employs advanced machine learning metrics; see Appendix A for more detailed information, namely here the most representative of the ROC curve and confusion matrix. The ROC curve, short for Receiver Operating Characteristic Curve, offers a visual depiction of the True Positive Rate (TPR), which signifies the probability of detection, juxtaposed with the False Positive Rate (FPR), indicative of false alarms across diverse threshold settings. Meanwhile, the confusion and error matrices provide a comprehensive visualisation of algorithmic performance in supervised learning. It delineates the percentages of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), offering critical insights into model accuracy and efficacy.

5.2.8 Data collection and analysis

5.2.8.1 Data and data collection methods

The data utilised in this validation endeavour comprise the following components: RDT Satellite data encompassing parameters such as Occurrence, Severity, and Altitude, NWP forecasts, Lightning and Radar Plan Position Indicator Product (PPI) data supplementing the meteorological datasets.

Discrete training, validation, and testing collections are meticulously assembled to facilitate the validation process by integrating high-resolution local NWP forecasts and RDT satellite images. This integration involves aligning the NWP grid with the higher-resolution satellite images and delineating the grid points within the RDT storm polygons. Interoperability between datasets is enhanced by standardising the data to a spatial resolution of .125 x .125 degrees.

Given the variance in temporal resolution (1 hour for NWP predictions compared to 15 minutes for RDT observations), an approach is devised to address this incongruity. A grid point is classified as convective if a storm observation is recorded during any of the four observation instances within the hour. This method enables constructing a binary training target function indicative of storm cell occurrence at a grid location within the hour.

Considering our focus on a 48-hour time horizon and the release of forecasts every 6 hours, disparate range forecasts valid for the same time frame are utilised for training, validation, and testing the model. This strategic adaptation accommodates the temporal dynamics inherent in the forecast data.

5.2.8.2 Analysis methods

Statistical analysis will be employed to evaluate the efficacy of the KAIROS solution 1 models, focusing specifically on the ML indicators, including the ROC curve and Confusion Matrix; see Appendix A for more information and indicators. Findings will be compared with observational data from convective

storm and lightning observations sourced from the RDT, lightning and Radar Plan Position Indicator Product (PPI) detection records.

5.2.9 Exercise planning and management

5.2.9.1 Activities

- Activity 1 - preparation of scenarios: Collect all weather data sources for the selected dates.
- Activity 2 – compute reference datasets: Prepare datasets with the target variables from the weather data sources.
- Activity 3 - compute model datasets: Extract dataset values for the selected dates.
- Activity 4 – data comparison and analysis of the results.
- Activity 5 – prepare the validation report.

5.2.9.2 Roles and responsibilities in the exercise

AI Methods will lead and develop the validation activities, Meteomatics will provide the NWP Forecasts products, RDTs are provided by MetSafe, lightning by MetSafe and Meteomatics and Cloud Top Height by the European Space Agency.

5.2.9.3 Time planning

ID	Q1 2024			Q2 2024			Q3 2024			Q4 2024		
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Activity 1												
Activity 2												
Activity 3												
Activity 4												
Activity 5												

Table 10: Detailed planning for EXE02 Phase 1

5.2.9.4 Identified risks and mitigation actions

Risks	Impact (1-low, 2-medium, 3-high)	Likelihood (1-low, 2-medium, 3-high)	Criticality (calculated based on likelihood and impact)	Mitigation actions

Risk #01: AI models do not provide results: AI models may not be able to capture the complex dynamics of certain weather phenomena.	High	Low	High	Previous research has shown that AI is able to generalize the complex behaviour of extreme weather events. AI algorithm development will look to past research to design AI models and algorithms
Risk #02: Data availability: Not having the necessary data will hinder attempts to develop AI models. It is important to have a large volume of relevant data in a timely manner to produce the models according to the project timeline.	High	Low	High	Required data sources have been identified. Most of the data will be available via the project partners (Meteomatics, MetSafe). The project will also need to purchase certain data sets, the sources of these additional data sets have been identified and initial cost estimates have been obtained.

5.3 Validation Exercise #03 plan. Local Scale Convection – Local forecast

5.3.1 Validation exercise description and scope

This exercise aims to validate the effectiveness of the KAIROS AI convection forecast/now-cast, specifically focusing on its impact on storm "now-casting" predictions at the local level and its support for tactical decision-making by Airport Operators. The AI convection forecast model utilises a sophisticated architecture, combining Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM). CNN and LSTM are artificial neural networks commonly used in pattern recognition and sequence prediction tasks.

As in EXE#02, this model is trained using Rapid-Development Thunderstorms (RDT), lightning strikes, Cloud Top height and Radar Plan Position Indicator Product (PPI).

The forecast temporal range spans 6 hours, and its geographical scope extends to the airport vicinities. The difference with the previous exercise is that the forecast employed as inputs for this model will include high-resolution local NWP. Through this exercise, our objective is to demonstrate the efficacy of the KAIROS AI convection forecast in refining storm predictions and supporting tactical decision-making for airport resources, including staff and runway configuration. This exercise will also be executed in 2 consecutive phases: the first phase will assess the model's accuracy, followed by activities to show its integration in real-time applications. Please note that the present version of the document will not focus on the stakeholders' benefits assessments; further phases will be completed in the future.

5.3.2 Stakeholders' expectations and benefit mechanisms addressed by the exercise.

Stakeholder	Involvement	Why it matters to the stakeholder
Airports	No active involvement in this initial VALP.	The throughput at airports can be negatively impacted by severe weather conditions. Having improved weather forecasts can allow for better planning of resources at airports.

5.3.3 Validation objectives

Objective	Explanation
Objective #01: Produce AI-based MET forecast	Apply artificial intelligence algorithms on available forecast and observation weather data to improve the prediction of several weather phenomena impacting aviation (convective weather).
Objective #02: Accuracy of prediction	Provide evidence that the prediction provided by the AI Convection Forecast tool produces accurate predictions of convective storms

Objective #03: Accuracy improvement	Provide evidence that the prediction provided by the AI Convection Forecast tool produces predictions of convective storms that are more accurate than existing tools available today
Objective #04: Forecast format	Check that the AI Convection Forecast is able to produce in real time predictions of convective storms adapted to the format required by the end-users
Objective #05: Number of predictions	Demonstrate that the AI Convection Forecast can produce forecasts at various leadtimes
Objective #06: Quality of predictions	Check that minimum data quality process is applied to the AI Convection Forecast input data and development process
Objective #07: Real-time performance monitoring	Validate that the AI Convection Forecast can perform real-time performance monitoring and alert in case the predictions produced have a level of confidence below the nominal standards
Objective #08: Post-ops analysis	Validate that the AI Convection Forecast is able to perform post-operational performance analysis

5.3.4 Validation scenarios

In Phase#01, the validation process will employ historical data from 2023 to formulate scenarios for analysis. Each day within these scenarios will undergo rigorous validation across multiple forecast releases within the KAIROS solution, ensuring a comprehensive assessment of its predictive capabilities. Validation scores will be meticulously derived for each storm characteristic, enabling detailed evaluation. In Phase#02, as a real-time analysis, the weather data will be from the day of the exercise execution. The area of interest for the KAIROS Solution 1 local model will cover the region of Istanbul and Zurich airports.

5.3.4.1 Reference scenario(s)

We should characterise the occurrence, severity and altitude of storms and the occurrence of lightning.

5.3.4.2 Solution scenario(s)

Solution scenarios are generated by applying the KAIROS solution, which enhances the prediction of storm occurrence, severity, altitude, and lightning using artificial intelligence techniques. Following this, it is necessary to establish indicators to evaluate the predictive capability of AI-enhanced weather forecasting against the reference scenario. Traditional machine learning metrics will be employed for this purpose, including Receiver Operating Characteristic (ROC) curve, Precision-Recall (PR) curve, and Confusion Matrix. Further details can be found in Appendix A.

5.3.5 Exercise validation assumptions

No additional assumptions apply to this exercise; see Section 4.4.

5.3.6 Limitations and impact on the level of significance

As in the previous exercise, when RDT is used as ground truth, the model results tend to overestimate storm sizes, which could imply inefficiency for ANSP activities such as planning ATC sector configuration or ATC tactical interventions. To better characterise storms, KAIROS models will be incorporated into the training phase storm observation from radar data.

5.3.7 Validation exercise platform/tool and validation technique

5.3.7.1 Validation exercise platform / tool characteristics

KAIROS solution 1 will be developed in Python, using the Keras library to implement the artificial recurrent neural network architecture. In-house computers will create every part of the AI model. A dedicated platform will also be devised to present the model's performance metrics. Appendix B illustrates a preliminary version of this tool.

5.3.7.2 Validation exercise Technique

The validation process employs advanced machine learning metrics; see Appendix A for more detailed information, namely here the most representative of the ROC curve and confusion matrix. The ROC curve, short for Receiver Operating Characteristic Curve, offers a visual depiction of the True Positive Rate (TPR), which signifies the probability of detection, juxtaposed with the False Positive Rate (FPR), indicative of false alarms across diverse threshold settings. Meanwhile, the confusion and error matrices provide a comprehensive visualisation of algorithmic performance in supervised learning. It delineates the percentages of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), offering critical insights into model accuracy and efficacy.

5.3.8 Data collection and analysis

5.3.8.1 Data and data collection methods

The data utilised in this validation endeavour comprise the following components: RDT Satellite data encompassing parameters such as Occurrence, Severity, and Altitude, NWP forecasts, Lightning and Radar Plan Position Indicator Product (PPI) data supplementing the meteorological datasets.

Discrete training, validation, and testing collections are meticulously assembled to facilitate the validation process by integrating high-resolution local NWP forecasts and radar image data. This integration involves aligning the NWP grid with the higher-resolution satellite images and delineating the grid points within the radar storm polygons.

Given the variance in temporal resolution (1 hour for NWP predictions compared to 15 minutes for RDT observations), an approach is devised to address this incongruity. A grid point is classified as convective if a storm observation is recorded during any of the four observation instances within the hour. This method enables the construction of a binary training target function indicative of storm cell occurrence at a grid location within the hour.

Considering our focus on a 6-hour time horizon and the release of forecasts every hour, disparate range forecasts valid for the same time frame are utilised for training, validation, and testing the model. This strategic adaptation accommodates the temporal dynamics inherent in the forecast data.

5.3.8.2 Analysis methods

Statistical analysis will be employed to evaluate the efficacy of the KAIROS solution 1 models, focusing specifically on the ML indicators, including the ROC curve and Confusion Matrix; see Appendix A for more information and indicators. Findings will be compared with observational data from convective storms and lightning observations sourced from the radar product and lightning detection records.

5.3.9 Exercise planning and management

5.3.9.1 Activities

- Activity 1 - preparation of scenarios: Collect all weather data sources for the selected dates.
- Activity 2 – compute reference datasets: Prepare datasets with the target variables from the weather data sources.
- Activity 3 - compute model datasets: Extract dataset values for the selected dates.
- Activity 4 – data comparison and analysis of the results.
- Activity 5 – prepare the validation report.

5.3.9.2 Roles and responsibilities in the exercise

AIMethods will lead and develop the validation activities, Meteomatics will provide the NWP Forecasts products, RDTs are provided by MetSafe, lightning by MetSafe and Meteomatics and Cloud Top Height by the European Space Agency.

5.3.9.3 Time planning

ID	Q1 2024			Q2 2024			Q3 2024			Q4 2024		
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Activity 1												
Activity 2												
Activity 3												
Activity 4												
Activity 5												

Table 11: Detailed planning for EXE03 Phase 1

5.3.9.4 Identified risks and mitigation actions

Risks	Impact	Likelihood	Criticality (calculated based on	Mitigation actions

	(1-low, 2-medium, 3-high)	(1-low, 2-medium, 3-high)	likelihood and impact)	
Risk #01: AI models do not provide results: AI models may not be able to capture the complex dynamics of certain weather phenomena.	High	Low	High	Previous research has shown that AI is able to generalize the complex behaviour of extreme weather events. AI algorithm development will look to past research to design AI models and algorithms
Risk #02: Data availability: Not having the necessary data will hinder attempts to develop AI models. It is important to have a large volume of relevant data in a timely manner to produce the models according to the project timeline.	High	Low	High	Required data sources have been identified. Most of the data will be available via the project partners (Meteomatics, MetSafe). The project will also need to purchase certain data sets, the sources of these additional data sets have been identified and initial cost estimates have been obtained.

6 References

6.1 Applicable documents

6.2 Reference documents

Appendix A AI performance metrics

The AI performance indicators analysis will be the first phase in the validation exercises.

Mean Absolute Error (MAE): the mean absolute difference between the real and the predicted values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_{real} - y_{predicted}|$$

Mean Absolute Percentage Error (MAPE): is just the MAE in percentage form.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_{real} - y_{predicted}}{y_{real}}\right|$$

Median Absolute Percentage Error (MDAPE): the median of the absolute percentage errors of the predictions with the real values

$$MDAPE = Median\left(\frac{|y_{real} - y_{predicted}|}{y_{real}}\right)$$

Root Mean Square Error (RMSE): calculated based on the residual for each data point.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\|y_{real} - y_{predicted}\|^2}{N}}$$

Confusion matrix in binary classification: a two-by-two table formed by counting the number of the four outcomes of a binary classifier. A binary classifier predicts all data instances of a test dataset as positive or negative. It is usually denoted as TP, FP, TN, and FN.

- True positive (TP): correct positive prediction
- False positive (FP): incorrect positive prediction
- True negative (TN): correct negative prediction
- False negative (FN): incorrect negative prediction

Accuracy: Expressed as $\frac{TP+TN}{TP+TN+FP+FN}$ it is a measure of how many predictions made by a model are correct out of the total predictions.

Precision: Expressed as $\frac{TP}{TP+FP}$, it indicates how many of the positively predicted instances were correct. It's useful when false positives are costly.

Recall or probability of detection: Sensitivity or True Positive Rate measures the model's ability to identify all positive instances correctly. Its equation is $\frac{TP}{TP+FN}$.

False alarm ratio: $\frac{FP}{TP+FP}$ this is the score used in ROC curves.

False alarm rate: $\frac{FP}{FP+TN}$

Critical success index (Threat score): $TS = \frac{TP}{TP+FP+FN}$

Co-funded by the European Union

Gilbert skill score: $ETS = \frac{TP-ar}{TP+FP+FN-ar}$ where $ar = \frac{(TP+FP)(TP+FN)}{n}$

Fraction Skill Score: for spatially distributed output. Skill scores can be calculated at different spatial resolutions. The Fractional Skill Score is a metric used in meteorology to evaluate the accuracy of predictions from numerical models, especially in the prediction of spatially distributed phenomena. This metric helps assess models' ability to predict meteorological events' structure, location and amplitude on spatial and temporal scales.

The FSS is calculated over several spatial scales to understand how model performance varies at different scales. A "scale" here refers to a specific area over which predictions and observations are averaged. For a given scale, the FSS is calculated as follows:

$$FSS = 1 - \frac{\sum_{i=1}^{N}\left(F_o^i - F_m^i\right)^2}{\sum_{i=1}^{N}\left(F_o^i\right)^2 + \sum_{i=1}^{N}\left(F_m^i\right)^2}$$

Where:

- F_o^i is the fraction of the observation in window i.
- F_m^i is the fraction of the model in window i.
- N is the total number of windows.

F1_score: Expressed as $\frac{2\,precision*recall}{precision+recall}$, it is a single metric that balances both precision and recall. F1_score provides a more comprehensive evaluation of a model's performance, even more so when an imbalance exists between the classes.

ROC curve: A ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. A ROC curve plots TPR vs. FPR at different classification thresholds. This curve plots two parameters:

- True Positive Rate $\frac{TP}{TP+FN}$
- False Positive Rate $\frac{FP}{FP+TN}$

Area Under the ROC Curve (AUC): measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus) from (0,0) to (1,1).

Intersertion Over Union (IoU): The Jaccard Index, also known as Intersertion Over Union (IoU), is a metric to evaluate the similarity between two data sets. It is especially useful in the context of images or data structured in two dimensions, where it is used to compare their degree of similarity. This index finds application in images with binary pixels, allowing us to measure how similar they are. The formula to calculate the IoU is:

$$IoU = \frac{|A \cap B|}{|A \cup B|}$$

Where:

- Intersection Area: It is the area in which the model predictions and the actual location of the phenomenon overlap.
- Union Area: It is the total area covered by the model predictions and the actual location of the phenomenon, that is, the sum of both areas minus the intersection area.

Moran index: this index is used to measure spatial autocorrelation, that is, how an attribute in a specific location is related to the same attribute in nearby locations. For example, let's imagine that we are analyzing the temperature in different cities. Nearby cities tend to have similar temperatures. If the Moran Index is calculated for this attribute, a high positive value would indicate that cities tend to have similar temperatures. If the value is close to zero, it means that there is no clear correlation in temperatures between nearby cities, suggesting a random distribution. A negative value would indicate that nearby cities tend to have very different temperatures (negative correlation). The key is how the value of the attribute at one location compares to the same value at nearby locations.

In the case of a map with binary predictions, this index could be applied to evaluate whether storms tend to cluster in certain areas or if they are distributed randomly. The Moran Index will analyse whether the cells with the detected phenomenon are close to each other or if they are randomly distributed. That is, the spatial autocorrelation of storm predictions within the image would be evaluated (whether the occurrence or absence of storms at one location is related to the occurrence or absence of storms at nearby locations).

Matthews Correlation Coefficient: is a metric used to measure the quality of binary classifications. It is useful in situations where classes are imbalanced, unlike other metrics such as precision or sensitivity, which can give a wrong impression of the model's performance in cases of class imbalance. This method is very convenient because, in storm prediction, it is more common that there are no storms than there are. Applying MCC in an analysis of cell-segmented images is appropriate, as it provides a global evaluation of the model's performance in its task of predicting specific events. This metric only considers the occurrence of an event in binary classification terms and does not incorporate spatial information about where that event occurs. It focuses exclusively on the quality of a model's predictions. The formula to calculate the MCC is:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

Appendix B KAIROS Model Analysis Dashboards

Figure 3 illustrates a preliminary version of the tool used to study the performance of the KAIROS models. Users can select various parameters within the lower section, such as date, threshold values, model layers, and flight levels. Concurrently, the central area exhibits three distinct subfigures: the comparison between predictions and actual outcomes, the ROC curve, and the confusion matrix. Figure 4 presents the observation (RDT data) for the selected date.

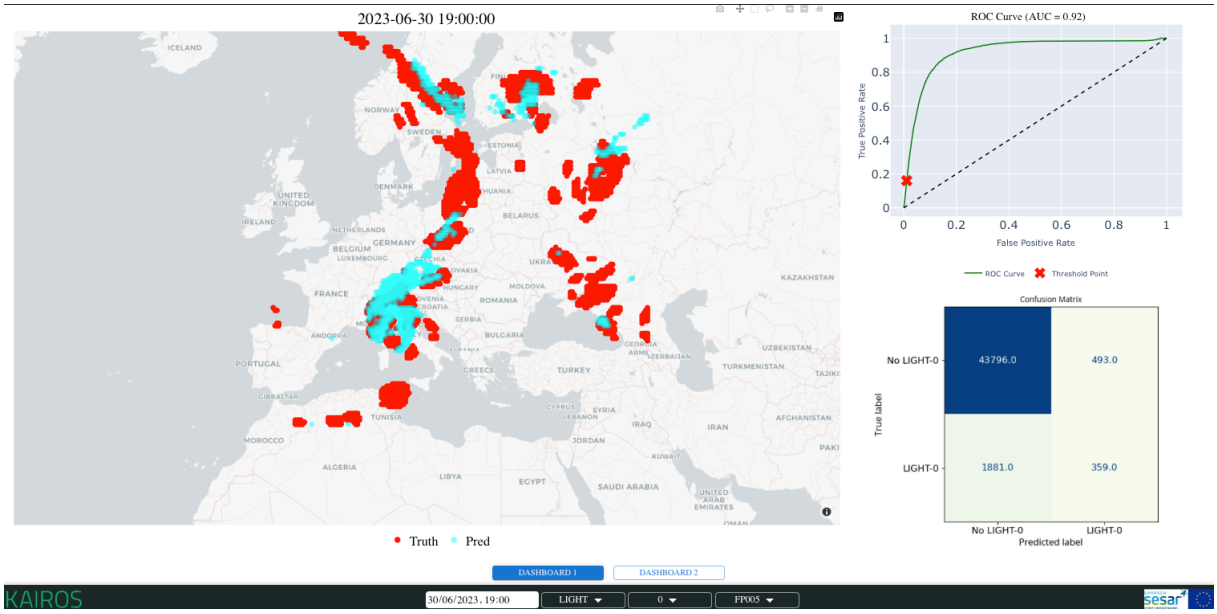

Figure 3. KAIROS model analysis tool Dashboard 1. Prediction vs Truth for a selected threshold.

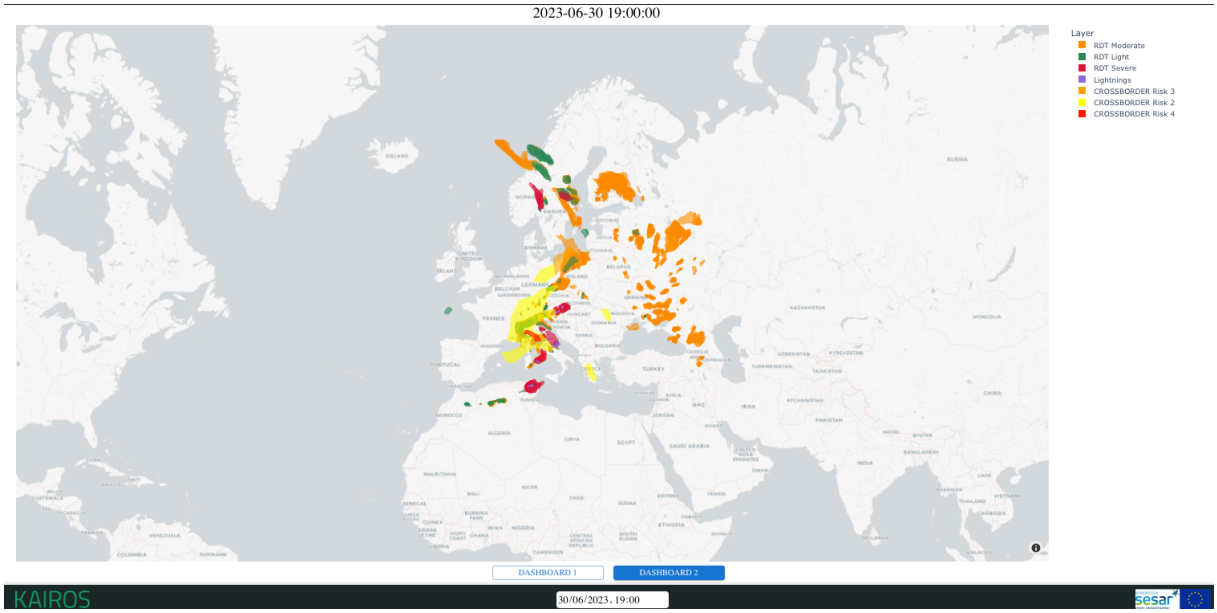

Figure 4 KAIROS model analysis tool Dashboard 2. RDT data (observation).