Scientific and Validation Report for the Extrapolated Imagery Processor of the NWC/GEO

NWC/CDOP4/MTG/GSA/SCI/VR/EXIM, Issue 1.0.0

31 March 2025

Applicable to GEO-EXIM v3.0 (NWC-045)

Prepared by GeoSphere Austria

For quick navigation go to: Table of contents

REPORT SIGNATURE TABLE

Function	Name	Signature	Date
Prepared by	P. Schmederer (GeoSphere Austria)		*31 March 2025*
Reviewed by	A. Jann, A. Wirth (both GeoSphere Austria)		*31 March 2025*
Authorised by	P. Rípodas (AEMET) *SAFNWC Project Manager*		*31 March 2025*

DOCUMENT CHANGE RECORD

Version	Date	Pages	CHANGE(S)
1.0.0	*31 March 2025*	56	Creation

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

The EUMETSAT's "Satellite Application Facilities" (SAFs) are dedicated centres of excellence for processing satellite data, and form an integral part of the distributed EUMETSAT Application Ground Segment (http://www.eumetsat.int). This documentation is provided by the SAF on Support to Nowcasting and Very Short-Range Forecasting, NWC SAF. The main objective of NWC SAF is to provide, further develop and maintain software packages to be used for Nowcasting applications of operational meteorological satellite data by National Meteorological Services. More information can be found at the NWC SAF webpage, http://www.nwc-saf.eumetsat.int.

1.1 SCOPE AND PURPOSE OF THE DOCUMENT

This document is a Validation Report for NWC/GEO Extrapolated Imagery (*EXIM*) Products (PGE16), based on the NWC/GEO release 2025, with the main scope on verifying continuity between MSG and MTG output. Moreover, some enhancements were tested as candidates for release in successor versions.

This document contains a description of the validation method and the corresponding results for the above-mentioned product.

1.2 DEFINITIONS, ACRONYMS AND ABBREVIATIONS

AMV	Atmospheric Motion Vector
ASII-NG	Automatic Satellite Image Interpretation – Next Generation
ASII-TF	Tropopause Folding sub-product of ASII-NG
BT	Brightness Temperature
BIAS	Bias
CDOP	Continuous Development and Operations Phase
CMA	Cloud Mask
CMIC	Cloud Microphysics
CRR	Convective Rainfall Rate
CRRPh	Convective Rainfall Rate from Cloud Physical Properties
CSI	Critical Success Index
CT	Cloud Type
CTTH	Cloud Top Temperature and Height
ECMWF	European Centre for Medium-Range Weather Forecasts
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellites
EXIM	Extrapolated Imagery
FAR	False Alarm Ratio

FCI	Flexible Combined Imager
FSS	Fraction Skill Score
HRVIS	High-resolution VISible
HRW	High-Resolution Winds
IFS	Integrated Forecasting System
IR	Infrared
MSG	Meteosat Second Generation
MTG-I1	Meteosat Third Generation – Imager 1
NWC	Nowcasting
NWP	Numerical Weather Prediction
PC	Precipitating Clouds
PCPh	Precipitating Clouds from Cloud Physical Properties
PGE	Product Generation Element
POFD	Probability of False Detection
POD	Probability of Detection
PSS	Peirce's Skill Score
RMSE	Root Mean Squared Error
SAF	Satellite Application Facility
SAFNWC	SAF to support Nowcasting and Very-Short-Range Forecasting
SEVIRI	Spinning Enhanced Visible and Infrared Imager
VIS	Visible
WV	Water Vapour

1.3 REFERENCES

1.3.1 Applicable Documents

The following documents, of the exact issue shown, form part of this document to the extent specified herein. Applicable documents are those referenced in the Contract or approved by the Approval Authority. They are referenced in this document in the form [AD.X].

For dated references, subsequent amendments to, or revisions of, any of these publications do not apply. For undated references, the current edition of the document referred applies.

Current documentation can be found at the NWC SAF Helpdesk web: http://www.nwc-saf.eumetsat.int.

Ref	Title	Code	Vers	Date
[AD.1]	Proposal for the Fourth Continuous Development and Operations Phase (CDOP 4) March 2022 – February 2027	/NWC/SAF/AEMET/MGT/CDOP4Proposal	1.0	12/03/2021
[AD.2]	Project Plan for the NWCSAF CDOP4 phase	NWC/CDOP4/SAF/AEMET/MGT/PP	3.0.0	21/10/2024
[AD.3]	Configuration Management Plan for the NWCSAF	NWC/CDOP4/SAF/AEMET/MGT/CMP	1.2.0	29/03/2024
[AD.4]	NWCSAF Product Requirement Document	NWC/CDOP4/SAF/AEMET/MGT/PRD	3.0.0	21/10/2024
[AD.5]	NWCSAF CDOP 4 Service Specifications	NWC/CDOP4/SAF/AEMET/MGT/SSD	1.0.0	31/10/2022
[AD.6]	System and Components Requirements Document for the NWC/GEO MTG-I day-1	NWC/CDOP2/MTG/AEMET/SW/SCRD	1.3.1	31/03/2025
[AD.7]	Architecture Design Document for the NWC/GEO MTG-I day-1	NWC/CDOP2/MTG/AEMET/SW/ACDD	1.3.0	31/03/2025
[AD.8]	Interface Control Document for Internal and External Interfaces of the NWC/GEO MTG-I day-1	NWC/CDOP2/MTG/AEMET/SW/ICD/1	1.4.0	31/03/2025
[AD.9]	Interface Control Document for the NWCLIB of the NWC/GEO MTG-I day-1	NWC/CDOP2/MTG/AEMET/SW/ICD/2	1.4.0	31/03/2025
[AD.10]	Data Output Format for the NWC/GEO MTG-I day-1	NWC/CDOP2/MTG/AEMET/SW/DOF	1.4.0	31/03/2025
[AD.11]	User Manual for the NWC/GEO: Software Part	NWC/CDOP3/MTG/AEMET/SW/UM	1.3.0	31/03/2025

Table 1: *List of Applicable Documents*

1.3.2 Reference Documents

The reference documents contain useful information related to the subject of the project. These reference documents complement the applicable ones, and can be looked up to enhance the information included in this document if it is desired. They are referenced in this document in the form [RD.X].

For dated references, subsequent amendments to, or revisions of, any of these publications do not apply. For undated references, the current edition of the document referred applies.

Current documentation can be found at the NWC SAF Helpdesk web: http://www.nwc-saf.eumetsat.int.

Ref	Title	Code	Vers	Date
[RD.1]	The Nowcasting SAF Glossary	NWC/CDOP4/SAF/AEMET/MGT/GLO	1.0	31/10/23
[RD.2]	User Manual for the Extrapolated Imagery Processor of the NWC/GEO: Science Part	NWC/CDOP3/MTG/ZAMG/SCI/UM/EXIM	1.2.0	31/03/25
[RD.3]	Algorithm Theoretical Basis Document for the Extrapolated Imagery Processor of the NWC/GEO MTG-I day-1	NWC/CDOP2/MTG/ZAMG/SCI/ATBD/EXIM	1.1.1	31/03/25
[RD.4]	Scientific and Validation Report for the Extrapolated Imagery Processor of the NWC/GEO	NWC/CDOP2/GEO/ZAMG/SCI/VR/EXIM	1.0	22/05/17
[RD.5]	Scientific and Validation Report for the Extrapolated Imagery Processor of the NWC/GEO	NWC/CDOP3/GEO/ZAMG/SCI/VR/EXIM	2.0.1	28/02/22
[RD.6]	Algorithm Theoretical Basis Document for the Precipitation Product Processors of the NWC/GEO	NWC/CDOP3/GEO/AEMET/SCI/ATBD/Precipitation	1.0.1	29/10/21

Table 2: *List of Referenced Documents*

2. GENERAL ASPECTS OF THE VALIDATION APPROACH

EXIM version 3.0 of the NWCSAF/GEO software v2025 has been the subject of this evaluation. The validation dataset for MSG-related aspects covers periods from each season of the year 2023. For the MTG-related aspects, only winter cases were available due to the timing constraints of when FCI was declared operational and when the first NWC/MTG release was expected. The set of extrapolated SEVIRI channels and NWCSAF products is listed in **Table 3**; a subset of the portfolio was selected to evaluate the functionality of *EXIM* for MTG-I1 in comparison to MSG.

Scores have been logged every (other) hour (see the range of scores in Chapter 2.5). For MSG forecasts, scores were recorded at lead times of +15, +30, +45 and +60 minutes. For MTG-I1, scores were calculated for the available steps: +10, +20, +30, +40, +50, +60 minutes. For the comparison of MTG-I1 with MSG, the spatial scores were calculated for the common lead times of 30 and 60 minutes. The study region is Europe, as shown in Figure 2.1.

Due to their nature, visible channels have only been considered at daytime.

Forecasts from the integrated forecasting system (IFS) at the European Centre for Medium Range Weather Forecasts (ECMWF) have been used as numerical weather prediction (NWP) input.

The version numbers of each product are listed in **Table 3**, high resolution winds (HRW) have been calculated with version 7.0.

Product (version)	**Abbreviation**	**Details**	**Evaluated aspects of chapter**			
			2.1	2.2	2.3	2.4
SEVIRI thermal infrared	*IR3.9*	3.9 µm		x		
	IR10.8	10.8 µm		x		x
SEVIRI visible	*VIS0.6*	0.6 µm	x			
	VIS0.8	0.8 µm	x			
High Resolution Visible	*HRVIS*		x			
Automatic Satellite Imagery Interpretation Next Generation - Tropopause folding sub-product(v3.0)	*ASII-TF*		x			
Convective Rainfall Rate from Cloud Physical Properties (v4.0)	*CRRPh*		x			x
Cloud Type (v5.0)	*CT*			x		x
Cloud Top Temperature and Height – sub-product “cloud top height in metres” (v5.0)	*CTTH alti*			x		x
Cloud Mask (v5.0)	*CMA*					x
Cloud Microphysics – cloud optical thickness sub-product (v3.0)	*CMIC cot*			x	x	
	CMIC phase					x

Precipitating Clouds from Cloud Physical Properties (v3.0)	*PCPh*					x

***Table 3**: SEVIRI channels and NWCSAF products evaluated in this validation report.*

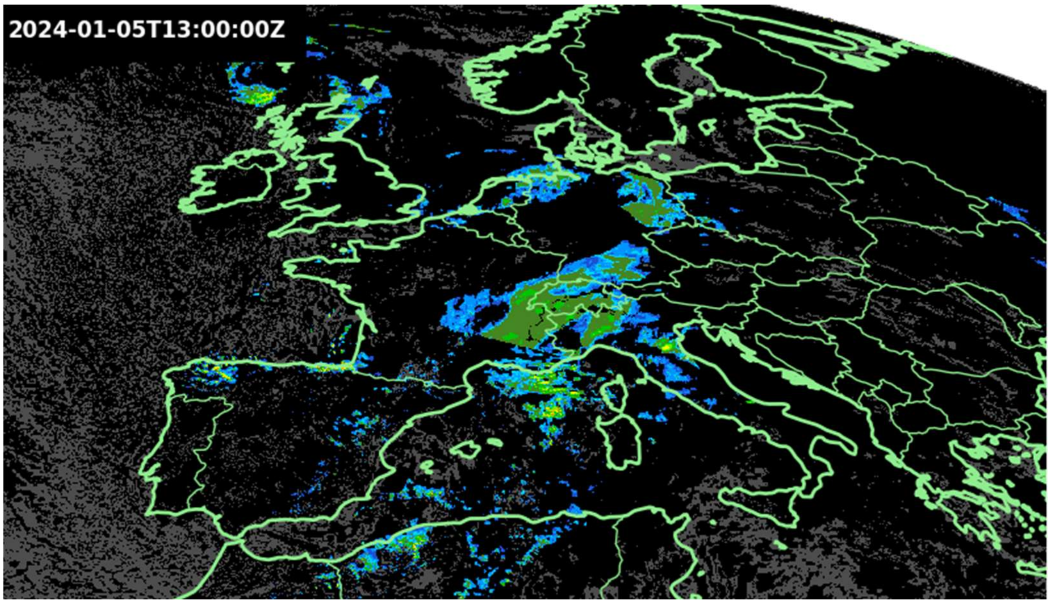

***Figure 2.1**: Geographical area over which the analyses have been performed.*

2.1 PRODUCT EVALUATION OF EXIM'S APPLICABILITY

This report covers the evaluation of new products and channels that may potentially be added to *EXIM's* portfolio in the version following the MTG day-1 release. It also evaluates whether already included products and channels can still be extrapolated with *EXIM* after they underwent a change in their algorithm. But above all, the document covers the validations about the continuity of the product suite from MSG to MTG-I1, see Chapter 2.4.

The following products are not yet included in the *EXIM*'s portfolio at MTG day-1: *HRVIS* (i.e. also the FCI VIS channels which by design have a higher resolution) and *ASII-TF*. The *EXIM* code under development has been improved and accelerated since the last evaluation (**[RD.5]**). Now the code would be capable to deal with the high resolution of the channel *HRVIS* and finish the forecasts of this product in due time (meaning in time before the next satellite images are available). Another code improvement to be released shortly is the implementation of a cosine correction, which is relevant for all visible channels, particularly during dusk/dawn. Therefore, besides *HRVIS* also two other MSG visible channels, 0.6 µm and 0.8 µm (VIS06 and VIS08), are evaluated in this validation. ASII-TF was also included in the evaluation as a potential new product to be added to the EXIM portfolio. Following a user request, cloud microphysics' "cloud optical thickness" (*CMIC cot*) has also been investigated as a potential addition to the portfolio (this addition could be easily taken on-board the first MTG release since the parameter has been in the internal *EXIM* portfolio for long but was not open to the users because evaluation results were not convincing. It turned out, however, that this was due to the choice of synoptically irrelevant thresholds in the first validation campaign).

Since *CRRPh* underwent algorithmic changes, it had to be re-evaluated to ensure that *EXIM* remains applicable to this product. For details on the modification to the algorithm, see **[RD.6]**.

All listed channels and products were extrapolated with *EXIM,* and their nowcasts were evaluated against persistence forecasts. According to the NWCSAF Product Requirements Table (PRT) ([AD.4]), the threshold accuracy of *EXIM* is described as "*on average better than persistence forecast*" and the target accuracy is "*always better than persistence forecast*".

The current default model version of *EXIM* was used in this evaluation. This setup is a one-layer scheme with lower and upper boundaries as defined in Chapter 2.2, thereunder named as "control" setup.

2.2 EVALUATION OF EXIM'S CTTH FILTER

In *EXIM* version v2021, the optional cloud top height checking (CTTH filter) was introduced as a user-configurable feature that allows filtering based on height. The user can define up to two layers. When enabled, the CTTH filter ensures that pixels are extrapolated only using AMVs from the one or two user-defined layers. The feature allows to distinguish between lower and upper weather phenomena or to focus on a specific layer by excluding AMVs from other heights. A comprehensive evaluation of the CTTH filter's performance, covering a range of products (*CMA*, *CMIC*, *CT*, *PC*, *PCPh*, *CRR*, *CRRPh*, *CTTH*) as well as satellite channels (*IR108*, *IR38*, *VIS08* and *VIS06*), was conducted in the last validation report (**[RD.5]**).

Some of the products needed to be re-evaluated for two reasons. First, the lowest boundary of the layers was chosen to be 900 hPa, which is sufficient for most of the products and channels. However, for *CTTH alti* (cloud top height in metres), *IR 10.8 µm* and *IR 3.8 µm*, this validation uses the value of 1100 hPa as lower boundary to ensure including all levels down to the very surface. Second, the validation of multi-categorical products was performed including all categories of the respective products. However, some of the categories describe surface types or do not have dedicated heights. This evaluation will exclude those categories in the validation of the CTTH filter. This is relevant for the two products *CT* and *CMIC*.

The five configurations used for the evaluation of the CTTH filter (referred to as "filter" below) are:

- "control setup": pixels of ***all*** heights are extrapolated with vectors from the heights 1100 – 100 hPa
- "low, no filter": pixels of ***all*** heights are extrapolated with vectors from the **low** layer (1100 – 500 hPa)
- "low, with filter": pixels from the **low** layer (1100 – 500 hPa) are extrapolated with vectors from the **low** layer (1100 – 500 hPa)
- "high, no filter": pixels from ***all*** heights are extrapolated with vectors from the **high** layer (500 – 100 hPa)
- "high, with filter": pixels from the **high** layer (500 – 100 hPa) are extrapolated with vectors from the **high** layer (500 – 100 hPa)
- "2-layer, with filter": pixels of **low** heights (1100 – 501 hPa) are extrapolated with vectors from **low** heights (1100 – 501 hPa) and pixels from the **high** layer (500 – 100 hPa) are extrapolated with vectors from the **high** layer (500 – 100 hPa)

2.3 EVALUATION OF THE BENEFIT OF INCLUDING WV VECTORS

For the *CMIC cot* product, this report evaluates whether water vapour (WV) atmospheric motion vectors (AMVs) should be included to the set of used AMVs. Currently, WV AMVs are only used for the extrapolation of the WV channels. All other products and channels by default use AMVs from channels: *IR108, IR120, VIS06, VIS08, HRVIS.*

For understanding the value of WV AMVs here, the configurations for the evaluation were:

- "all AMVs, excluding WV": same as "control" setup from Chapter 2.2; AMVs from: IR108, IR120, VIS06, VIS08, HRVIS
- "all AMVs, including WV": similar to "control, without WV AMVs" with **additional AMVs** from: WV062, WV073
- "only WV": similar to "control, without WV AMVs" but **solely AMVs** from: WV062, WV073

2.4 Comparison MTG-I1 vs MSG

With MTG-I1 now operational and sending data, the NWCSAF product *EXIM* was evaluated to ensure that there are no major changes in functionality and quality compared to the precursor Meteosat satellite, MSG-3. A selection of products and imagery from the entire portfolio was chosen to represent the different types: satellite imagery (*IR108*), categorical products (*CMA, CT, CMIC phase*), and continuous products (*CTTH alti, CRRPh, PCPh*). This evaluation report covers two approaches: one spatial and one grid-point-based.

For the spatial verification, windows with the same geographical extension (though with different numbers of pixels) were used to calculate the fraction skill score (FSS). Specifically, 75×75 pixels of an MTG-I1 product/image cover the same area as 50×50 pixels of an MSG product/image. For the point-based scores, each satellite was evaluated against its respective persistence run to compare their behaviour at the default lead times of 30 and 60 minutes.

Additionally, the one-layer scheme with lower and upper boundaries, as defined in Chapter 2.2 and referred to as the "control" setup, was used for the comparison. The scores for the MSG-3 satellite were calculated only for lead times of 30 and 60 minutes, whereas the scores for the MTG-I1 satellite were calculated for all lead times.

2.5 Methodology

For the analysis, dichotomous scores such as POD, POFD, FAR, CSI and PSS (definitions are provided below), along with the spatial score FSS, are used. Forecasts, including *EXIM* and persistence forecasts, are validated against satellite images or derived product at verification time. Depending on the verification question, either the different *EXIM* versions are compared with each other, or *EXIM* forecasts are compared with their respective persistence forecasts. The spatial verification directly compares EXIM forecasts based on the two satellites, MSG and MTG-I1, against each other (chapter 3.4). Note that the evaluations presented in Chapter 3.1 to 3.3 were calculated with MSG data, while Chapter 3.4 consists of comparisons between both MSG and MTG-I1. As stated above, the threshold accuracy is: "*on average better than persistence*". Persistence is based on the initial state at lead time +0 minutes which is considered the forecast for subsequent lead times. Both persistence and EXIM forecasts are validated against the observed truth and compared with each other.

Dichotomous scores categorise the data, in this case, pixel-wise, in "yes, the event will happen" and "no, the event won't happen". For most of the products / satellite images, thresholds have been specified by separating "yes" and "no", in the sense of "the value *is* greater than/equal to/smaller than the threshold" and "the value *is not* greater than/equal to/smaller than the threshold", respectively. Multi-category products were classified similarly for each category and a multi-category skill score was computed, as described below.

To calculate dichotomous scores, one starts with a contingency table (**Table 4**) that shows the frequency of forecasts and occurrences in the domain and their joint distribution of hits (a), false alarms (b), misses (c) and correct negatives (d). Such a contingency table was produced for every lead time (+15, +30, +45, +60 minutes) and resulted in a set of scores for each lead time and time step averaging over the whole domain.

For the grid-point-based comparison of MTG-I1 versus MSG, the verification tool *harp* was used to calculate the (grid-) point scores. *harp* is a set of R packages developed within the ACCORD consortium and is designed for analysing and verifying NWP data. For this evaluation, a routine has been written to bring the satellite data in the correct input format for *harpPoint*. For the grid-point-based calculations of the scores and the plotting, *harpPoint* and *harpVis* routines were applied, respectively.

The used scores are defined as described in the following. For a more comprehensive description see https://www.cawcr.gov.au/projects/verification/ or Jolliffe and Stephenson[1] (2012).

	observed			
forecast		**yes**	**no**	total
	yes	a hits	b false alarms	a + b forecast yes
	no	c misses	d correct negatives	d + d forecast no
	total	a + c observed yes	b + d observed no	n = a + b + c + d total

Table 4*: Contingency table showing frequency of "yes" and "no" forecasts and occurrences.*

2.5.1 POD

The probability of detection (POD) answers the question what fraction of observed "yes" events was correctly forecasted. This score is good for rare events but ignores false alarms.

$$POD = \frac{a}{a+c} \qquad \text{Range } [0,1].$$

[1] Jolliffe, I.T., and D.B. Stephenson, 2012: Forecast Verification: A Practitioner's Guide in Atmospheric Science. 2nd Edition. Wiley and Sons Ltd, 274 pp.

2.5.2 FAR

The false alarm ratio (FAR) answers the question what fraction of predicted "yes" events actually did not occur. This score ignores misses and is good for rare events.

It is defined as:

$$FAR = \frac{b}{a+b} \qquad \text{Range } [0,1].$$

2.5.3 POFD

The probability of false detection (POFD), also called false alarm rate (F) answers the question what fraction of observed "no" events were incorrectly forecasted as "yes". This score ignores misses. Note that it can be artificially improved by issuing fewer "yes" forecasts to reduce the number of false alarms.

$$POFD = \frac{b}{b+d} \qquad \text{Range } [0,1].$$

2.5.4 CSI

The critical success index (CSI), also denoted threat score (TS) answers the question how well forecasted "yes" events correspond to observed "yes" events. It can be thought of accuracy when correct negatives have been removed.

$$CSI = \frac{a}{a+\ b+c} \qquad \text{Range } [0,1].$$

2.5.5 PSS

The Peirce Skill Score (PSS), also known as true skill statistic (TSS) or Hanssen and Kuipers discriminant (HK), answers the question how well the forecast separates the "yes" events from the "no" events. PSS is a measure of skill obtained by the difference between POD and POFD and is defined as:

$$PSS = POD - POFD = \frac{ad - bc}{(a+c)(b+d)} \qquad \text{Range } [-1,1].$$

If PSS is greater than zero, the number of hits exceeds the number of false alarms and the forecast has some skill.

	Observed category					
Forecasted category	i, j	1	2	…	K	Total
	1	$n(F_1, O_1)$	$n(F_1, O_2)$	…	$n(F_1, O_K)$	$N(F_1)$
	2	$n(F_2, O_1)$	$n(F_2, O_2)$	…	$n(F_2, O_K)$	$N(F_2)$
	…	…	…	…	…	…
	K	$n(F_K, O_1)$	$n(F_K, O_2)$	…	$n(F_K, O_K)$	$N(F_K)$

	Total	$N(O_1)$	$N(O_2)$	…	$N(O_K)$	

***Table 5**: Multi-category contingency table showing frequency of forecasts and occurrences in various bins.*

Two products are multi-category forecasts and therefore need to be treated slightly different. The products are:

- Cloud Type (*CT*), with 15 different cloud types;
- Cloud Microphysics' sub-product cloud phase (*CMIC phase*), with the five categories "liquid", "ice", "mixed", "cloud-free", and "un-defined".

For those, a multi-categorical variant of the PSS was used (PSS_{mc}). This method also starts with a contingency table (**Table 5**) showing the frequency of forecasts and occurrences of each bin. PSS_{mc} is defined as:

$$PSS_{mc} = \frac{\frac{1}{N}\sum_{i=1}^{K} n(F_i O_i) - \frac{1}{N^2}\sum_{i=1}^{K} N(F_i)N(O_i)}{1 - \frac{1}{N^2}\sum_{i=1}^{K}(N(O_i))^2} \qquad \text{Range } [-1,1].$$

2.5.6 BIAS

The bias or mean error describes the average forecast error.

$$BIAS = \frac{1}{N}\sum_{1=1}^{N}(F_i - O_i) \qquad \text{Range } [-\infty, \infty].$$

2.5.7 RMSE

The root mean squared error (RMSE) shows the average magnitude of the forecast error.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2} \qquad \text{Range } [0, \infty].$$

2.5.8 FSS

The fraction skill score compares the fractional coverage of two fields of grid-box events in spatially defined windows. These event frequencies are used directly to compute a Fractions Brier Scores, a version of the (half) Brier Score. The FSS is defined as follows:

$$FSS = 1 - \frac{\frac{1}{N}\sum_N \left(P_{mtg} - P_{msg}\right)^2}{\frac{1}{N}\left(\sum_N {P_{mtg}}^2 + \sum_N {P_{msg}}^2\right)} \qquad \text{Range } [0,1].$$

To define a spatial window, the grid-points are set to 75×75 pixels for MTG and 50×50 pixels for MSG, which covers the same geographical area.

The value of FSS above which the forecasts are considered useful (better than random) is given by FSS_{useful}:

$$FSS_{useful} = 0.5 - \frac{f_{msg}}{2} \quad ,$$

where f_{msg} is the domain average of the MSG (“observed”) fraction. FSS signals “skilful” in case where $FSS > FSS_{useful}$.

3. RESULTS

Results for all four evaluation aspects

- Comparison of EXIM forecasts against persistence,
- Evaluation of CTTH Filter,
- Evaluation of WV AMV inclusion, and
- comparison of MTG-I1 against MSG

are discussed in this chapter.

3.1 COMPARISON OF EXIM FORECASTS AGAINST PERSISTENCE

The following validation covers products that are evaluated for potential addition to the *EXIM* portfolio, including *HRVIS*, *ASII-TF*, and *CMIC cot*, as well as current portfolio members that have undergone algorithmic changes (either in the extrapolation or in the product itself), such as *VIS06*, *VIS08*, and *CRRPh*.

3.1.1 Visible Channels: HRVIS

The extrapolation of the high resolution visible channel (*HRVIS*) became available for this evaluation. An upgrade of the algorithm allows faster calculations and makes the inclusion of this channel possible. *HRVIS* is evaluated for reflectivity values within the range of 9, 15, 20, 30, 40, and 50%. Due to the nature of this channel, only day time is considered.

EXIM forecasts for *HRVIS* show improved scores compared to persistence, as illustrated in Figure 3.1. POD is enhanced for all thresholds and lead times. While the improvement is almost negligible, with values of the order 0.01 for small thresholds and lead times, it increases with both lead time and threshold, reaching an improvement of up to 0.08. FAR also increases for thresholds greater than 9, with the improvement becoming more pronounced as the threshold values increase, peaking with a reduction in FAR by 0.06. Overall, PSS indicates an improvement of *EXIM* forecasts compared to persistence for thresholds greater than 9.

While CSI (see Figure 3.2) shows only a slight improvement for smaller values, the enhancement becomes more pronounced for higher values, with an improvement of up to 0.6. As lead times increase, a slight bias towards over-predicting is introduced, as observed in the performance diagram of Figure 3.2.

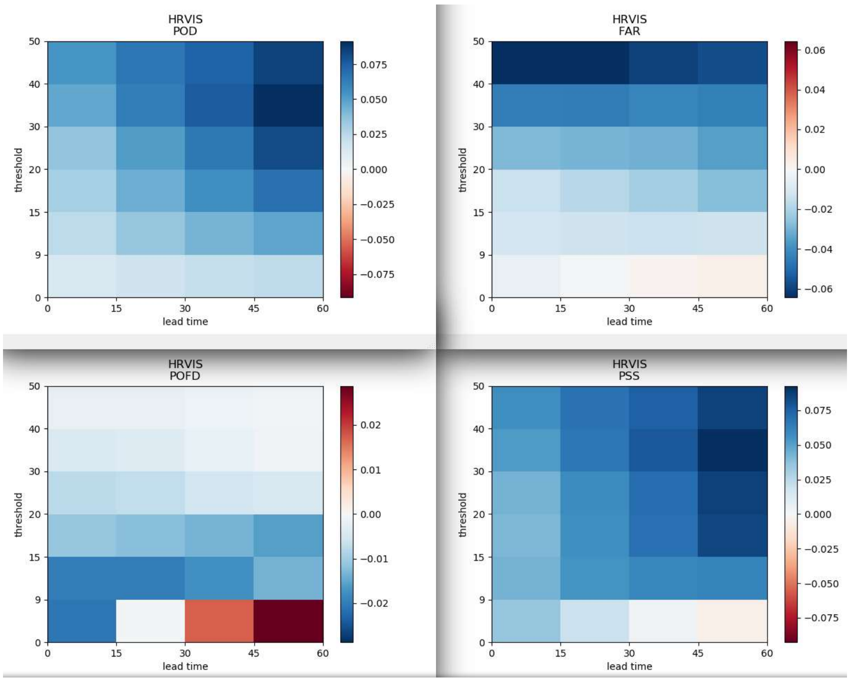

***Figure 3.1**: Change of scores due to EXIM forecasts compared to persistence, blue being an improvement, red a degradation. The scores are top left: POD, top right: FAR, bottom left: POFD, bottom right: PSS. Each of the figures displays the lead times (15, 30, 45, 60 min) on the x-axis and thresholds (reflectivity) on the y-axis (9, 15, 20, 30, 40, 50).*

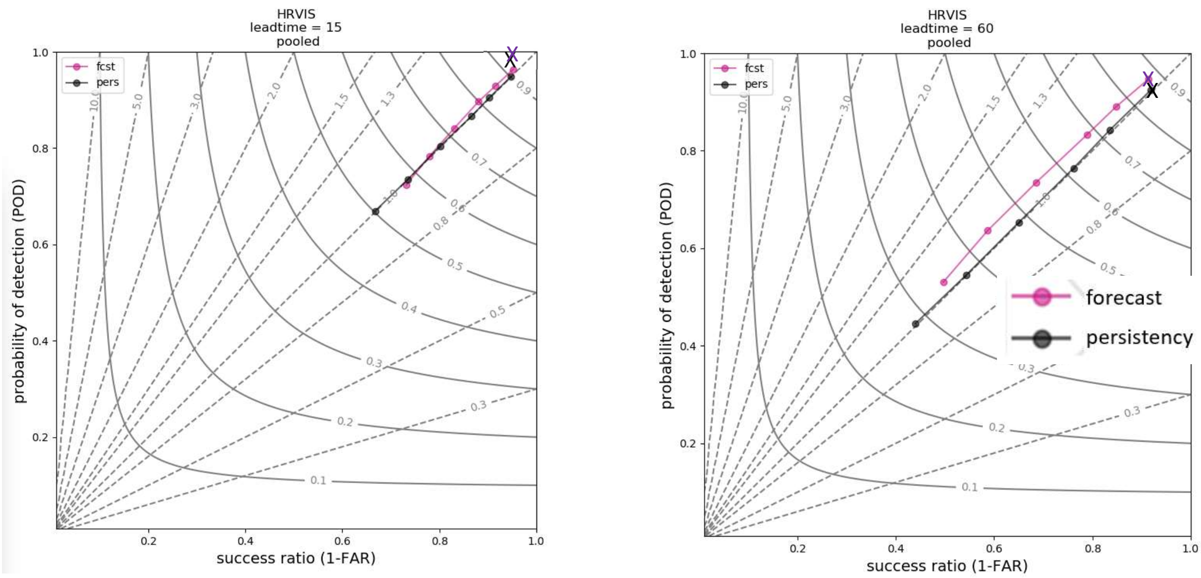

***Figure 3.2:** Performance diagram for HRVIS purple: EXIM forecast, black: persistence; left: lead time 15 min, right: lead time 60 min. The dots represent the different thresholds. The smallest threshold is marked with X for orientation.*

3.1.2 Visible Channels: VIS 0.6 µm

The visible channel with a wavelength of *0.6 µm (VIS06)* is evaluated for reflectivity values in the range of 9, 15, 20, 30, 40, 50. Only daytime data has been considered due to the nature of this channel. The new algorithm incorporates a cosine correction.

Similar to *HRVIS, EXIM* forecasts improve scores compared to persistence for the *VIS06* channel, as shown in **Figure 3.3**. POD is enhanced for all thresholds and lead times. While the improvement is modest for small thresholds (+ 0.02), it increases with higher lead times and thresholds, reaching a gain of 0.1. FAR is slightly reduced for the lowest thresholds, while it is improves for higher thresholds, with the improvement reaching a maximum of 0.08. While POFD shows only slight differences, the overall change of PSS is positive, with scores improving by up to 0.1, especially for a lead time of 60 minutes.

Even though a small bias is introduced in the forecasts compared to persistence, with a tendency towards too many "yes" events (**Figure 3.4**), the improvement in CSI outweighs the degradation caused by the bias.

There is no seasonal variability observed in the scores (not shown).

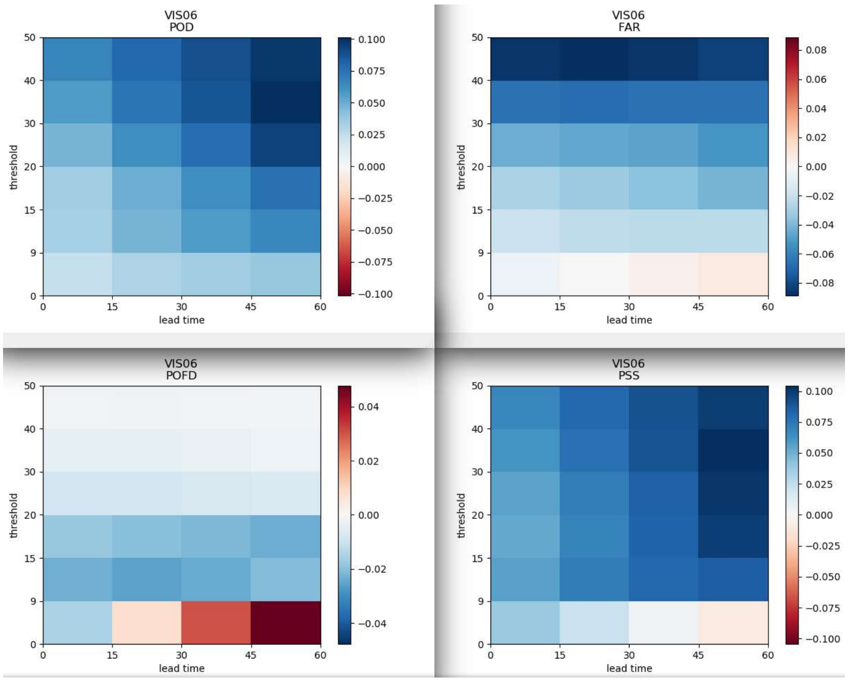

***Figure 3.3**: Like Figure 3.1, but for VIS06.*

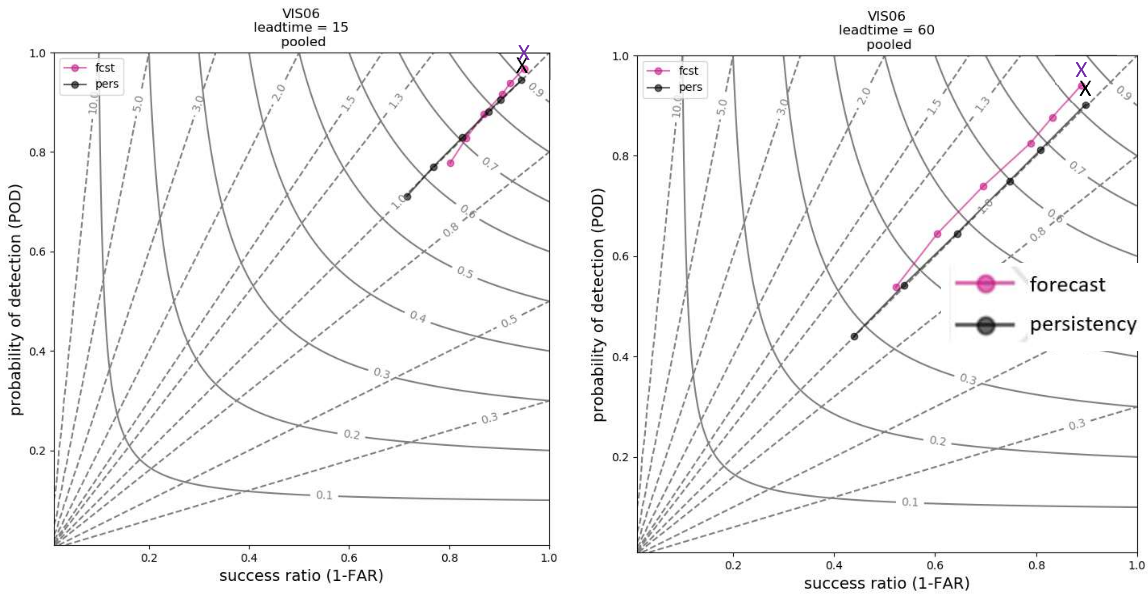

***Figure 3.4**: Performance diagram like Figure 3.2, but for VIS06.*

3.1.3 Visible channels: VIS 0.8 µm

The visible channel with a wavelength of *0.8 µm (VIS08)* is evaluated for reflectivity values in the range of 9, 15, 20, 30, 40, 50. Similar to *VIS06*, *VIS08* has been evaluated only at the daytime.

The results of *VIS08* are similar to those of the other two visible channels described above. *EXIM* forecasts perform better than persistence, as shown in **Figure 3.5**. POD is improved in *EXIM* forecasts compared to persistence, with the improvement becoming more pronounced as thresholds and lead times increase, reaching an increase of nearly 0.08. Meanwhile, FAR is reduced for thresholds above 9, with the improvement becoming more significant as the thresholds rise. The changes in POFD are minor but positive, with a maximum value of 0.03 for the lowest three thresholds; for the higher thresholds, the change is negligible. PSS outlines a positive impact of *EXIM* forecasts compared to persistence with negligible changes for threshold 9 but distinct improvement reaching up to 0.08 depending on lead time and threshold.

From the performance diagram, in **Figure 3.6**, one can see that there is an overall improvement of the forecasts. The improvement of CSI outweighs the slight bias, which is introduced by the *EXIM* forecast.

No seasonal variability was observed in the validation (not shown).

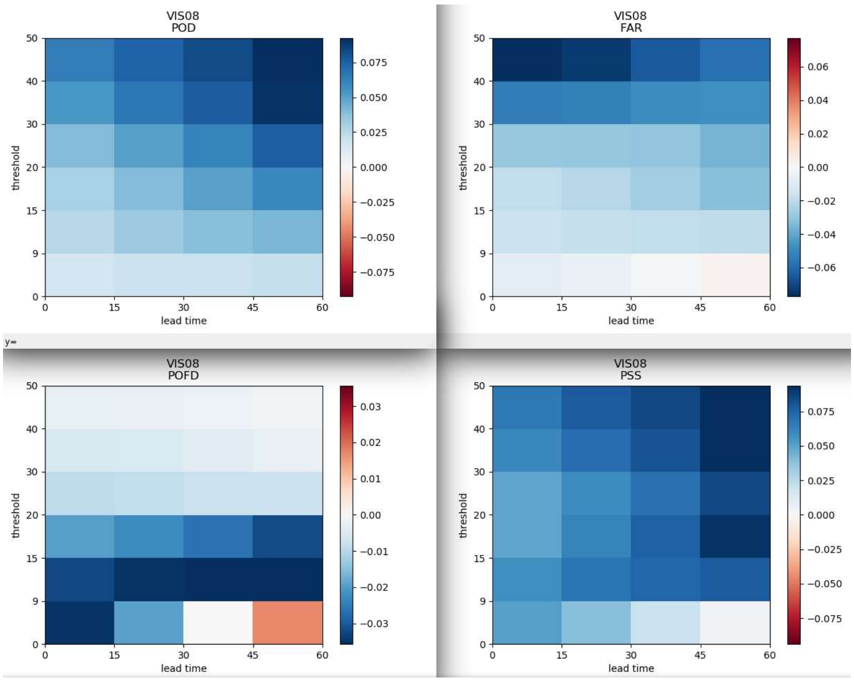

***Figure 3.5**: Like Figure 3.1, but for VIS08.*

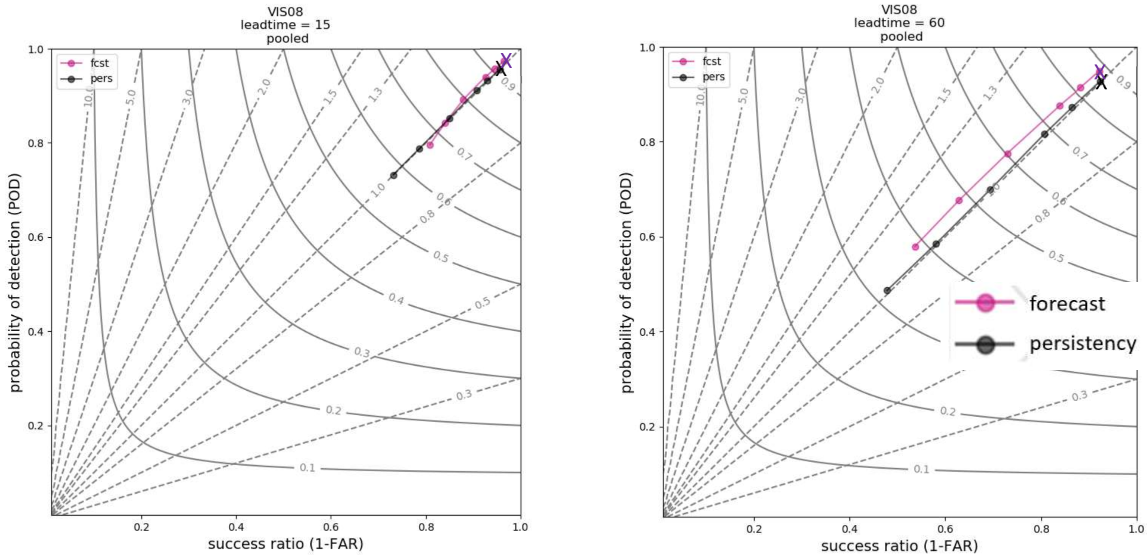

***Figure 3.6**: Performance diagram like Figure 3.2, but for VIS08.*

3.1.4 ASII-TF: Automatic Satellite Image Interpretation - Next Generation Tropopause Folding

Tropopause folding is one of the two potential candidates for inclusion in *EXIM*'s portfolio. The product describes the likelihood for tropopause folding, expressed as a percentage ranging from 0 to 100.

The *EXIM* forecasts for *ASII-TF* do not show an improvement in scores compared to persistence, as shown in **Figure 3.7**. The POD of persistence is better than the *ASII-TF* forecasts by *EXIM* for all thresholds and lead times, although the actual difference is very small for thresholds around 10 %. The degradation increases as the thresholds rise, reaching a maximum of 0.06. Also FAR degrades in the forecasts compared to persistence. With almost no change in POFD, the effect remains more negative than positive, leading to a degradation in PSS for all thresholds and values.

The ROC curve (not shown) and the performance diagram (**Figure 3.8**) display the same results. For lower thresholds, the difference in POFD, FAR and PSS are less pronounced compared to higher thresholds. The POFD is slightly increased in the *EXIM* forecasts compared to persistence, except for low thresholds, where there is no difference in greater thresholds. The CSI is degraded, showing almost no change for the lowest thresholds, and a decrease of about 0.06 for the highest thresholds.

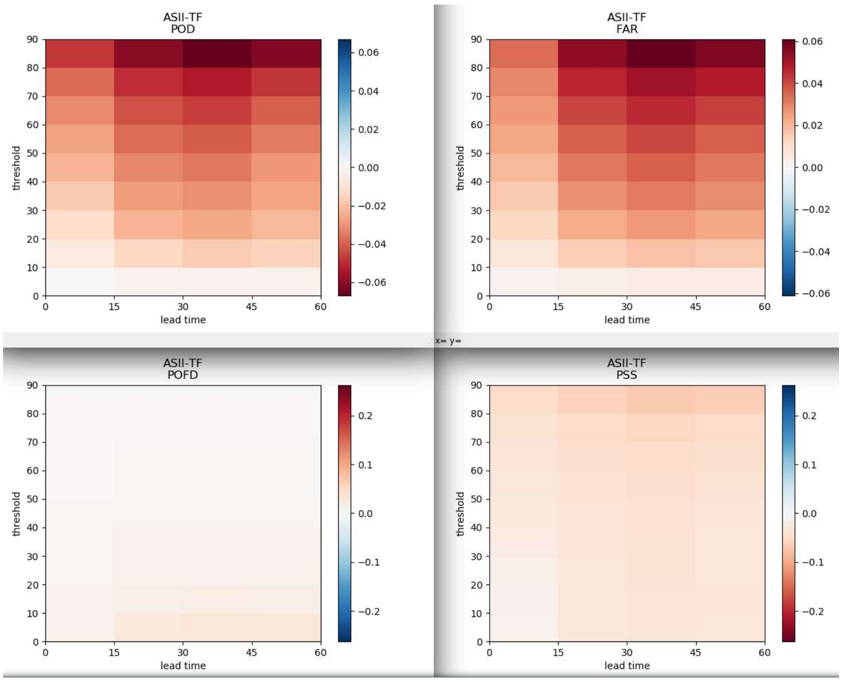

Figure 3.7: *Like Figure 3.1, but for ASII-TF.*

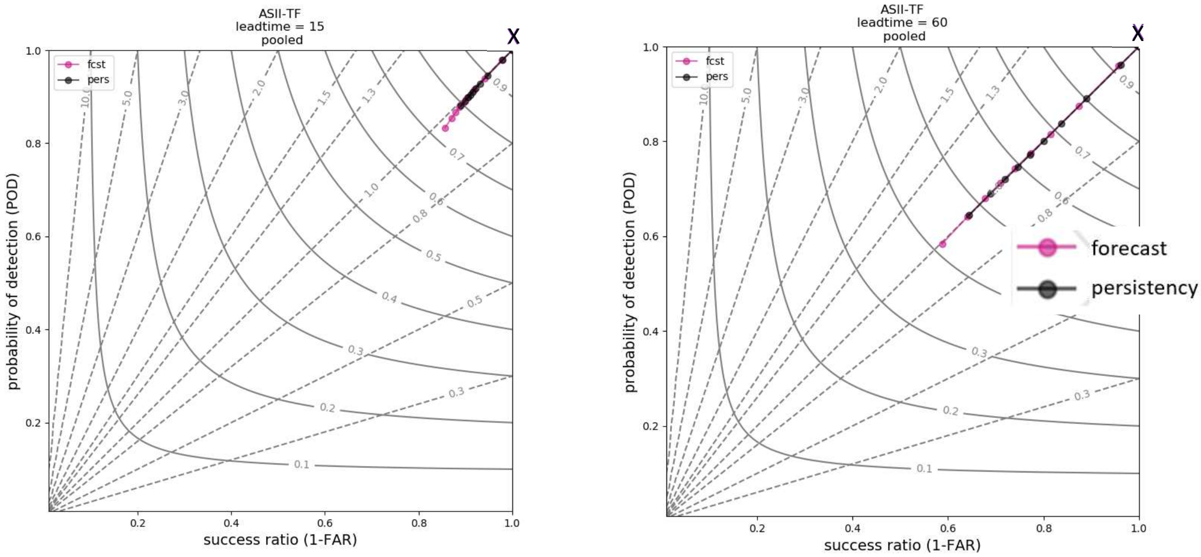

__Figure 3.8__: Performance diagram like in __Figure 3.2__, but for ASII-TF.

3.1.5 CMIC cot: Cloud Microphysics – cloud optical thickness

The Cloud Microphysics – Cloud Optical Thickness (*CMIC cot*) is the second of the two products potentially to be added to *EXIM*'s portfolio. *CMIC cot* provides cloud optical thickness and is defined as:

$$\text{cmic\_cot} = \text{scale\_factor} * \text{counts} + \text{offset},$$

with scale_factor being 0.01 and offset being 0. Values start from 1, unit = "1"; the thresholds for dichotomous scores are: 1, 2, 3, …, 17, 18.

EXIM's forecast for *CMIC cot* improves the scores compared to persistence, as shown in **Figure 3.9**. The POD is improved in *EXIM* forecasts compared to persistence for all lead times and thresholds. While the improvement is relatively small (of the order 0.02), it becomes more pronounced as thresholds increase, reaching an improvement of up to 0.1. The positive change in POD is more pronounced for shorter lead times and weakens slightly with longer lead times. FAR is reduced in *EXIM* forecasts similarly to POD, with the most pronounced improvement of about 0.08 for short lead times and high thresholds, and the smallest improvement of about 0.02 for long lead times and low thresholds. The POFD doesn't change significantly between the two forecasts; however, for small lead times and low thresholds, *EXIM* forecasts slightly improve POFD by about 0.03 compared to persistence. PSS shows an improvement for all lead times and thresholds, with the greatest improvement of more than 0.1 observed for high thresholds and shorter lead times.

With increasing thresholds, CSI is continuously improved from tiny changes, reaching to 1 and 0.5 for lead time 15 and 60 minutes, respectively (see **Figure 3.10**). The differences are relatively small. However, since adding this product to the extrapolated products was a user request, the main conclusion is that the scores do not contradict the inclusion to the *EXIM* portfolio.

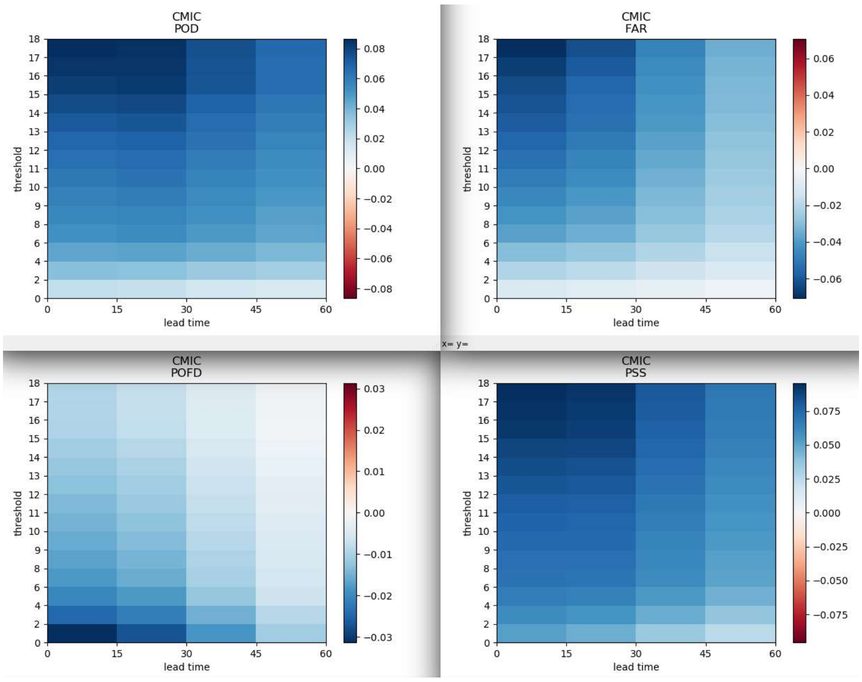

***Figure 3.9**: Like Figure 3.1, but for CMIC cot.*

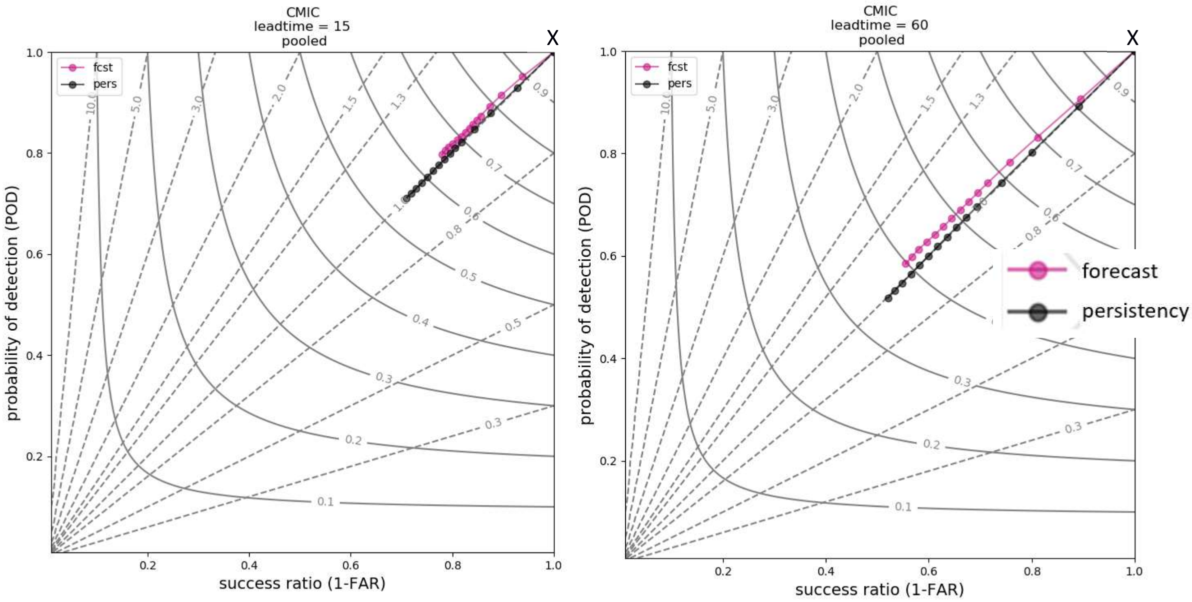

***Figure 3.10**: Performance diagram like in Figure 3.2, but for CMIC cot.*

3.1.6 CRRPh: Convective Rainfall Rate from Cloud Physical properties

The Convective Rainfall Rate based on Cloud Physical properties *(CRRPh)* underwent some algorithmic changes and is re-validated in this report to ensure it can still be accurately extrapolated with *EXIM*. The product estimates rain rates from convective clouds through cloud top microphysical information such as cloud top effective radius and cloud optical thickness. The thresholds for the dichotomous scores are: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 21 mm/h.

Forecasts of *EXIM* for *CRRPh* continue to improve POD compared to persistence. *EXIM* performs better than persistence as shown in **Figure 3.11**. POD improves by about 0.05 and 0.1. The improvement in FAR is of a similar magnitude to that of POD, with a slight gradient showing stronger improvement for shorter lead times. POFD shows little change when comparing *EXIM* forecasts and persistence forecasts. PSS improves for all lead times and thresholds below 11 mm/h. High values occur too infrequently to provide representative statistics.

From the performance diagram (**Figure 3.12**), it can be seen that CSI improves by about 0.1 for *EXIM* forecasts compared to persistence at the exemplarily shown threshold of 1 and 5. For greater thresholds, the improvement decreases gradually (not shown).

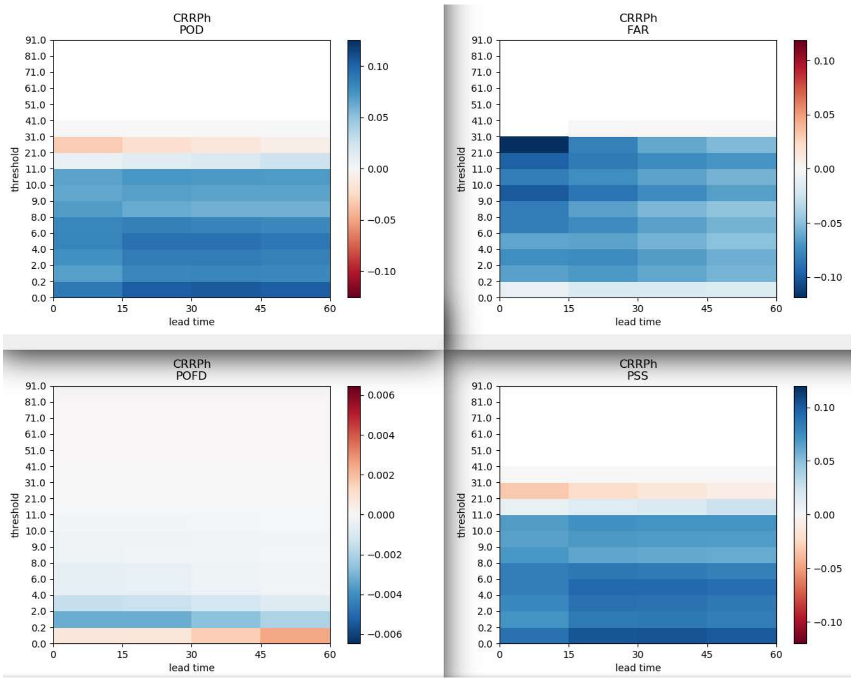

***Figure 3.11**: Like Figure 3.1, but for CRRPh.*

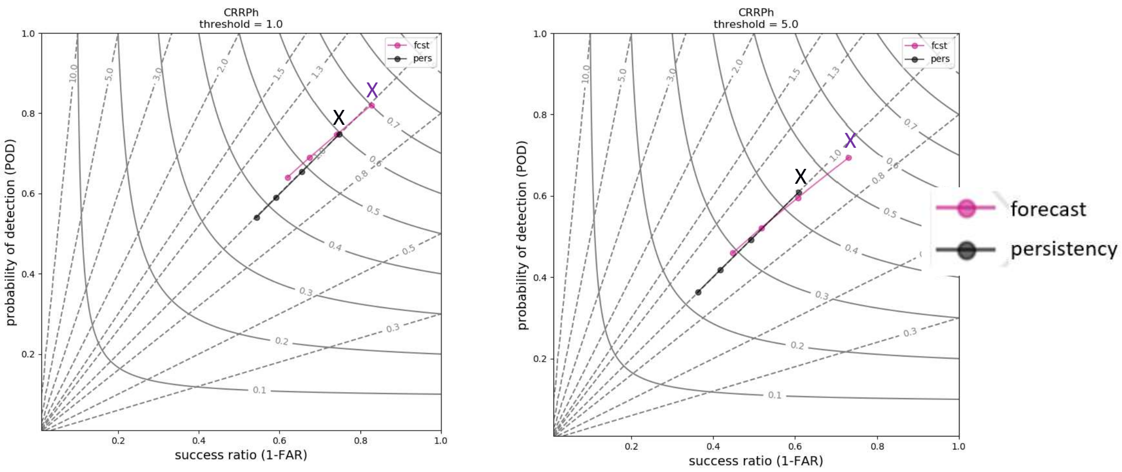

***Figure 3.12**: Like performance diagram in Figure 3.2, but for CRRPh. However, the dots are representing lead times instead of thresholds. The shortest lead time is marked with X for orientation.*

3.2 EVALUATION OF CTTH FILTER

The CTTH filter has been introduced to *EXIM* as a new feature in NWC/GEO v2021, allowing a filtering of pixels by height. Up to two layers can be defined by the user, separating e.g. a low-level from a high-level or selecting a specific layer of interest. With the filter in use, pixels will be extrapolated by *EXIM* using the closest set of atmospheric motion vectors (AMVs) stemming from the same layer as the pixels do. The idea is to avoid an extrapolation of pixels with AMVs stemming from a completely different height. The filter has already been validated in [RD.4]. The multi-categorical products *CT* and *CMIC phase* are re-evaluated considering this time only relevant categories and dropping the ones that will not be extrapolated. In addition, *CTTH alti*, *IR108* and *IR38* will be re-evaluated in this report including pressure levels down to the surface.

For the two layers, the lower layer spans from 1100 to 500 hPa and the upper layer extends from 501 to 100 hPa. The no-filter versions, as well as the “control”, span from 1100 to 100 hPa. The following analysis will compare the model setups outlined in Chapter 2.2. Each product will be addressed individually in the following sections.

3.2.1 Thermal Channels: Infrared 10.8 µm

The infrared channel with wavelength of *10.8 µm (IR108)* is evaluated in the range of 230 to 280 K brightness temperature (BT). *IR108* is strongly height-dependent. Low BTs typically correlate with high altitudes, while high BTs generally indicate low altitudes, However, there are exceptions, such as semi-transparent clouds.

The results from this evaluation agree well with the previous evaluation ([RD.4]). With the filter in use, scores vary strongly with BTs, as shown in ***Figure 3.13***. The low-level pixel extrapolation ("low, with filter") begins to show better skill than the high-level pixel extrapolation ("high, with filter") for BTs greater 270 K (the second highest of the investigated BT). There is an overall decrease of scores with increasing lead times. "low, with filter" has high POFD values for low BTs reaching up to 1 and "high, with filter" has low POD values for high BTs. Both setups with filter never perform best. This might be caused by the fact that their sample size is smaller.

The two setups that extrapolate all pixels with either just low ("low, no filter") or just high ("high, no filter") AMVs have a skill close to the control setup's, each of them having only a slightly reduced skill in the layers their AMVs are not associated with.

The forecasts from the "2-layer, with filter" setup perform very well compared to the 'control' forecasts for low BTs, but slightly lose skill for greater BTs.

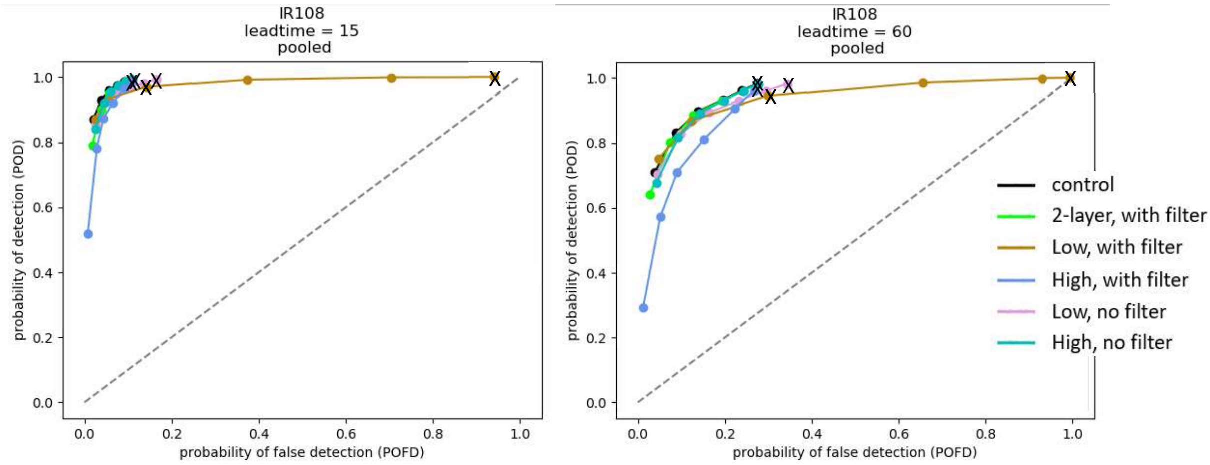

Figure 3.13: *ROC curve for IR108 with the x-axis being POFD and y-axis POD. The different versions are "control" (black), "2-layer, with filter" (green), "low, with filter" (ochre), "high, with filter" (blue), "low, no filter" (rose), "high, no filter" (turquoise). The top left corner in the diagram is the best. Left: lead time 15 minutes; right: lead time 60 minutes.*

3.2.2 Thermal Channels: Infrared 3.8 µm

The infrared channel with wavelength of *3.8 µm (IR38)* is evaluated in the range of 250 to 280 K BT. The behaviour is similar to that of *IR108*. The height dependency is evident when examining the scores in **Figure 3.14**.

Like *IR108*, the two setups, "low, with filter" and "high, with filter", perform worst for BTs associated with heights outside their respective layers. "Low, with filter" has greater POFD values for small BTs than the other versions, while "high, with filter" has lower POD than the other versions. The low POD values for the "high, with filter" version are most pronounced during the winter period. The high POFD values for the "low, with filter" version and the lowest BTs are more pronounced during the summer and autumn periods (not shown). In their best scenarios – at lead time 15 minutes and for low (high) BTs – forecasts from "high, with filter" ("low, with filter") achieve similar skill as the "control".

Similar to *IR108*, the "control" setup is not outperformed by any of the other setups overall.

The forecasts of the "2-layer, with filter" setup are almost of the same skill as those of the "control" setup.

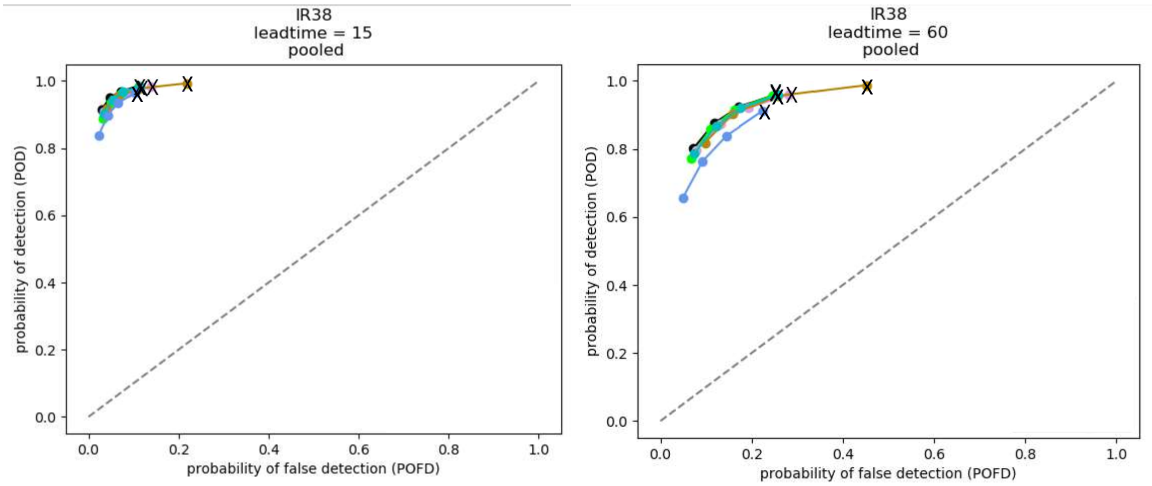

Figure 3.14: *ROC curve like* **Figure 3.13***, but for IR38.*

3.2.3 CTTH alti: Cloud Top Temperature and Height - altitude

The Cloud Top Temperature and Height Altitude (*CTTH alti)* is evaluated over a height range from 1000 m to 10 000 m, with 1000 m intervals.

The results of this evaluation confirm the results of the previous one ([RD.4]). The *EXIM* forecasts for *CTTH alti* show a strong response to the use of the filter, as clearly demonstrated in **Figure 3.15**. For forecasts using the "low, with filter" setup, there is minimal skill in POD and PSS for thresholds above 5000 m, with both scores dropping to 0.4. No pixels are extrapolated by *EXIM* for the "low, with filter" setup at thresholds greater than 6000 m. POD, POFD and PSS are zero. Conversely, the POFD of the "high, with filter" setup only starts to decrease below 1 at thresholds above 5000 m, and the PSS starts to show skill at thresholds greater than 7000 m. This result is expected and confirms the effectiveness of the filtering process.

Although the forecasts of the "low, with filter" and "high, with filter" setups never perform best, the forecasts of the "2-layer, with filter" setup compete very well with the "control" setup. PSS and FAR of the "2-layer, with filter" setup have a similar skill as the "control" setup, which also confirms the expectations regarding the filtering process.

There was a quality drop visible in the previous evaluation ([RD.4]) which used 900 hPa as lowest boundary definition in POFD and PSS of "low, with filter" and "2-layer, with filter" at the lowest altitude threshold. This cannot be observed in this evaluation for the "low, with filter" setup. The "2-layer, with filter" setup still experiences a slight degradation in PSS for the lowest thresholds, due to a small increase in POFD.

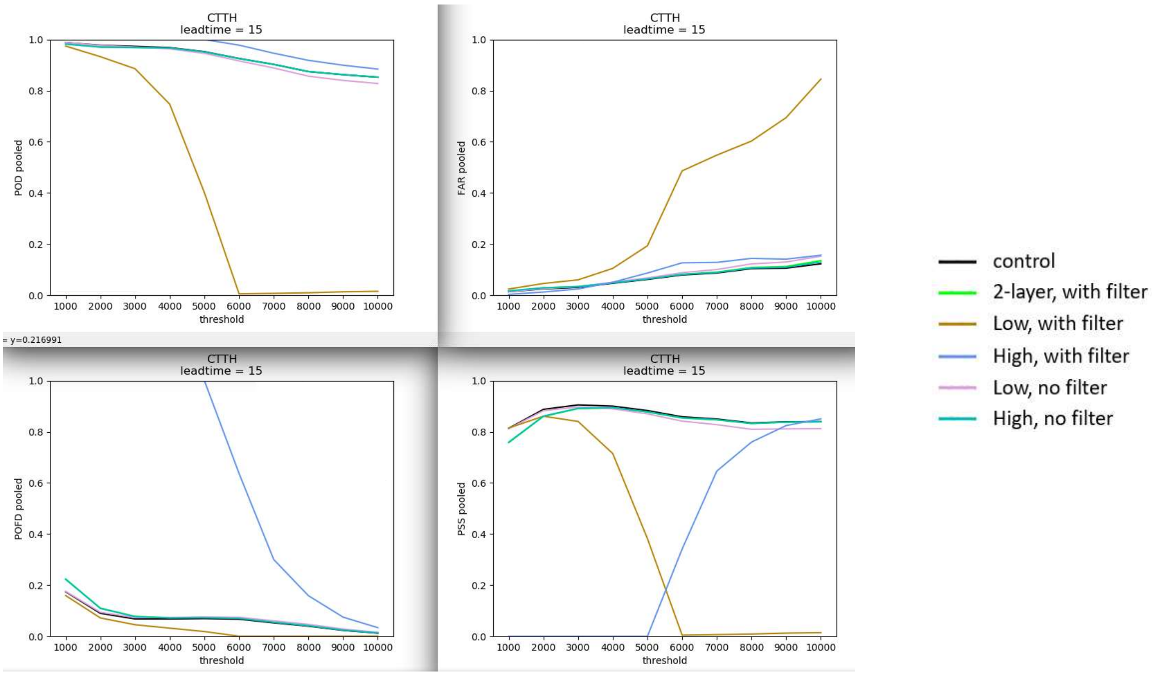

Figure 3.15: *Four scores for CTTH alti listed per category at lead time 15 min. Top left: POD, top right: FAR, Bottom left: POFD, bottom right: PSS.*

3.2.4 CT: Cloud Type

The product Cloud Type (*CT*) is one of the two multi-categorical products in this report. The categories are: cloud-free land (1), cloud-free sea (2), snow over land (3), sea ice (4), very low clouds (5), low clouds (6), mid-level clouds (7), high opaque clouds (8), very high opaque clouds (9), fractional clouds (10), high semi-transparent thin clouds (11), high semi-transparent moderately thick clouds (12), high semi-transparent thick clouds (13), high semi-transparent above low or medium clouds (14), high semi-transparent above snow ice (15).

Contrary to the previous evaluation ([RD.4]), this report considers only clouds with assigned heights but no surface types (1, 2, 3 and 4) or cloud type 10 (which doesn't have a dedicated height) for the multi-categorical Peirce skill score PSS_{mc}.

The outcome of this evaluation is different to the previous one ([RD.4]) as five types are excluded. As shown in **Figure 3.16**, the CTTH filter has a positive effect on the forecasts of the clouds types when only clouds are considered that can actually be extrapolated. The "2-layer, with filter" setup has the highest PSS_{mc} score, the "high, with filter" and the "low, with filter" setups perform second best with a reduction of 0.1 compared to the "2-layer, with filter" setup. The "control", the "high, no filter" and the "low, no filter" setups perform worst with a PSS value lower by roughly 0.2 compared to the "2-layer, with filter" setup. The differing results (compared to the the previous report) can be explained by the types that are not included in this evaluation. The complete absence of those types in the filter-versions, compared to their inclusion in the "control", the "high, no filter" and the "low, no filter" setups, makes a significant difference in the results of these two evaluations.

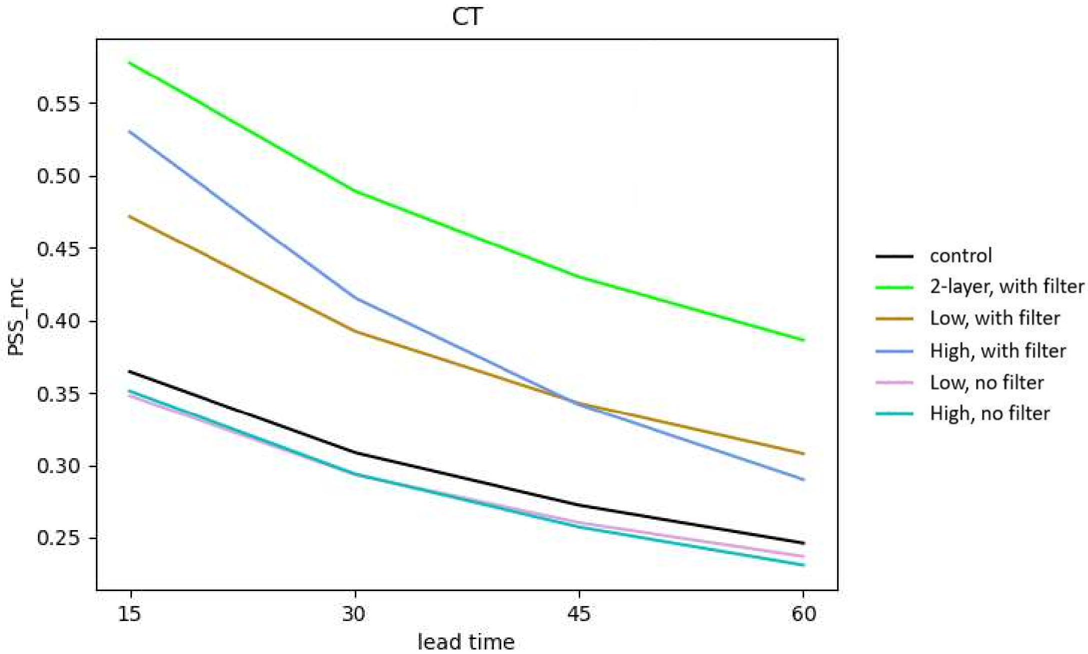

Figure 3.16: *Peirce skill score for multi-categorical vs lead times for CT. "Control" (black), "2-layer filter" (green), "low, with filter" (ochre), "high, with filter" (blue), "low, no filter" (rose), "high, no filter" (blue).*

3.2.5 CMIC phase: Cloud Microphysics - cloud phase

The Cloud Microphysics - Phase (*CMIC phase*) is the second of the two multi-categorical products. The categories are liquid (1), ice (2), mixed (3), cloud-free (4), and un-defined (5). The multi-categorical score PSS_{mc} has been computed in addition to the other scores. While the previous validation included all types, this one only includes 1, 2 and 3 ([RD.4]).

The ranking of the versions in this evaluation (see **Figure 3.17**) differs from the previous evaluation. Excluding cloud-free and un-defined categories results in the "2-layer with filter" setup being the best option. Next, with a degradation of just over 0.1, are the "control", the "low, no filter", the "high, no filter" setups. For a lead time of 15 minutes, the "low, with filter" setup compares well with these, but its performance declines more sharply as lead time increases.

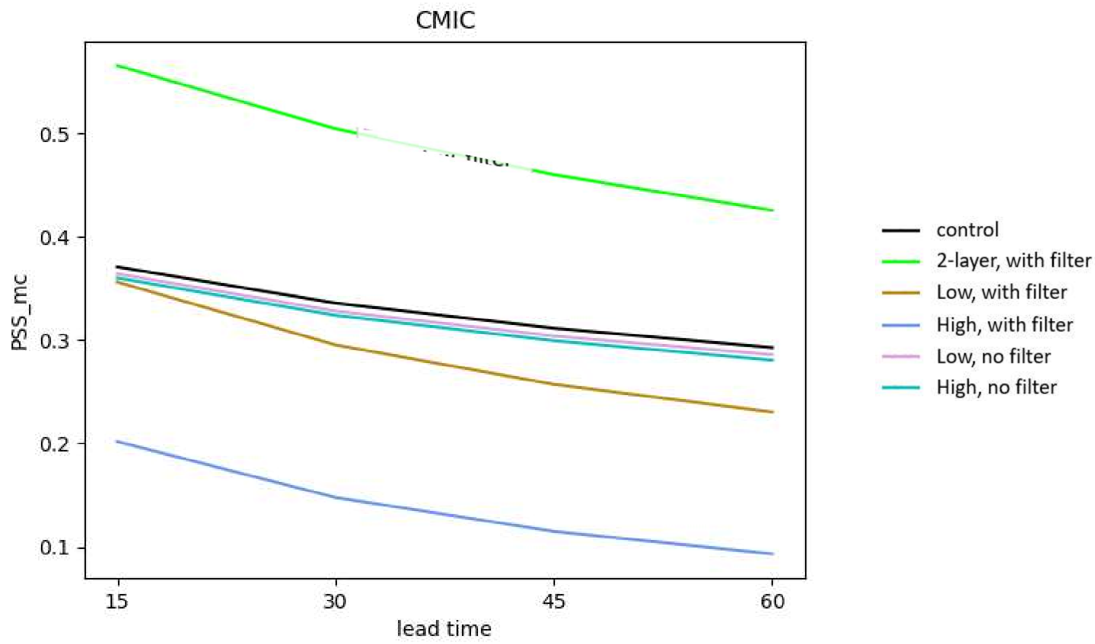

Figure 3.17: *Like Figure 3.16, but for CMIC phase.*

3.3 Evaluation of WV AMV inclusion

This chapter discusses the results of the evaluation regarding whether including water vapour (WV) atmospheric motion vectors (AMVs) should be considered for the *CMIC cot* product. *CMIC cot* is verified, as described in Chapter 3.1.5, for the values ranging from 1 to 18.

The results show that for *CMIC cot*, the best-performing set of AMVs is the "all but WV AMVs" setup. As shown in **Figure 3.18**, although there are differences among the three versions, the "all AMVs, excluding WV" setup performs best. There is only a slight degradation for the "all AMVs, excluding WV" setup, while using the "WV only" setup results in the lowest scores. This is true for all four scores displayed in **Figure 3.18**. In terms of POD, FAR and PSS, there is no difference for small thresholds, while for POFD, this holds true for greater thresholds. In this report, only the 15-minutes lead time is shown, but longer lead times exhibit a similar pattern, albeit with slightly lower scores.

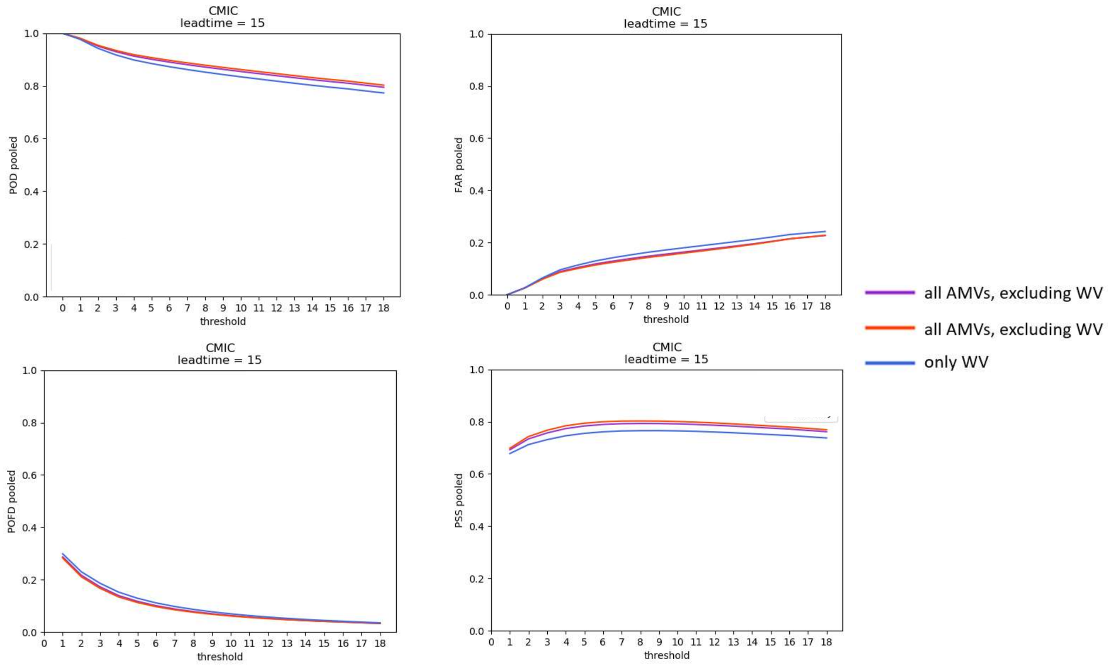

***Figure 3.18:** Like **Figure 3.15**, but for CMIC cot. The versions are "all AMVs, including WV" (purple), "all AMVs, excluding WV" (red), "only WV" (blue).*

3.4 COMPARISON OF MTG-I1 VS. MSG

3.4.1 Thermal Channels: Infrared 10.8 µm

The infrared channel with a wavelength of 10 µm (*IR108*) was evaluated in the range of 230 to 280 K brightness temperature (BT). The threshold scores were derived by identifying values smaller than the specified thresholds. The IR 10.8 µm channel was used as the representative for the satellite images.

Figure 3.19 displays the spatial scores, FSS and $FSS - FSS_{useful}$, for *IR108*. The FSS is nearly 1, indicating a very high agreement, except for the 210 K threshold, where FSS is low (0.5), which should, however, not be over-interpreted due to very few cases at this low threshold. The figure on the right shows the useful FSS subtracted from the FSS. As long as the FSS is greater than the useful FSS, the agreement is not solely due to chance. After subtracting FSS_{useful}, the agreement between the two channels remains high (ranging from 0.5 to 0.35). Only at very warm thresholds (270 and 280 K) and the coldest threshold (210 K) does the agreement decrease slightly to 0.2 and 0, respectively. The primary focus for *EXIM* is how the scores evolve with increasing lead time compared to lead time 0 minutes, as the latter reflects only the initial state without any interaction from *EXIM*. The slight decrease in scores is negligible, indicating that *EXIM* does not alternate the results and works similarly for both satellites.

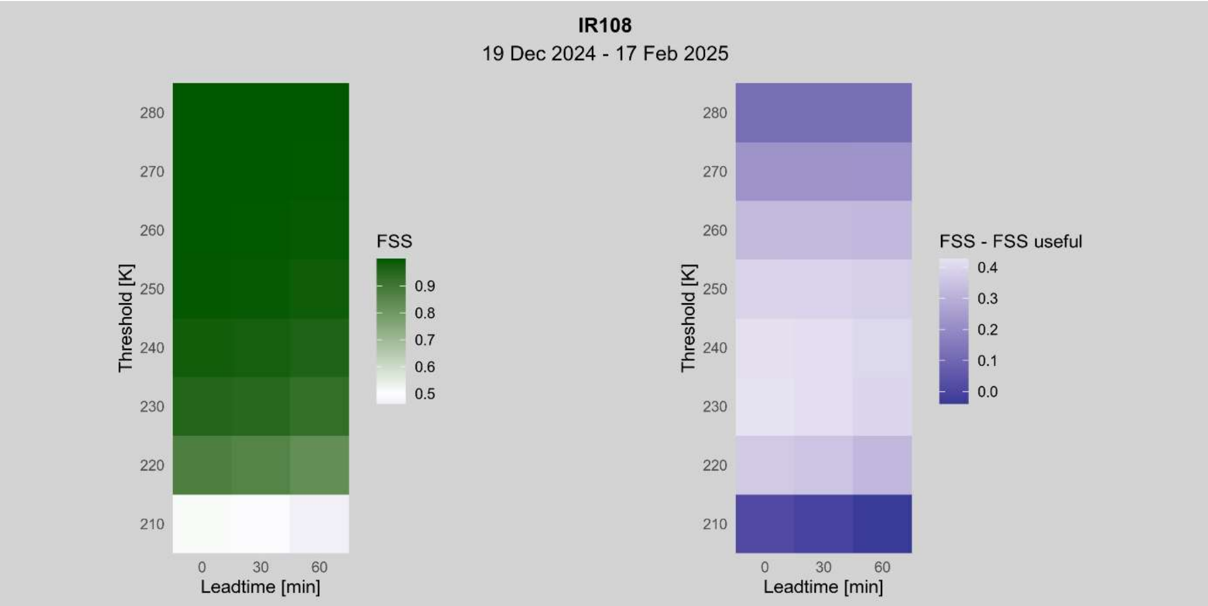

***Figure 3.19**: FSS (left) and $FSS - FSS_{useful}$ (right) for IR 10.8 µm channel: On the x-axis, lead times in minutes are displayed, and on the y-axis, the different thresholds in Kelvin are shown. The results represent the average scores of a winter period. Window size is set to 75×75 for MTG-I1 and 50×50 pixels for MSG, to cover the same geographical area.*

Considering the grid-point-based scores, where lead times at 30 and 60 minutes are comparable, it is evident that RMSE is reduced in *EXIM* forecasts compared to their respective persistence forecasts. However, the improvement for the MTG-I1 satellite is greater (a reduction of 2.5) than for MSG (a reduction of 0.5). The same trend is observed for POD (not shown) and PSS (see **Figure 3.20**), where *EXIM* shows higher values than persistence. Again, the improvement is more pronounced for MTG-I1 than for MSG. This is particularly noticeable for smaller thresholds (BTs of 230, 240, 250 K) compared to higher (warmer) thresholds. For the highest thresholds (270, 280 K), there is almost no difference between the satellites and the forecasts compared to persistence.

FAR and POFD (both not shown) are reduced in *EXIM* forecasts compared to their respective persistence forecasts. The colder the threshold, the more pronounced the difference between *EXIM* and persistence forecasts. Additionally, the improvement is greater for MTG-I1 than for MSG. For example, the maximum reduction in FAR is about 0.8 for MTG-I1 versus 0.4 for MSG at lead time 30 minutes, and 0.5 versus 0.3 at lead time 60 minutes, respectively.

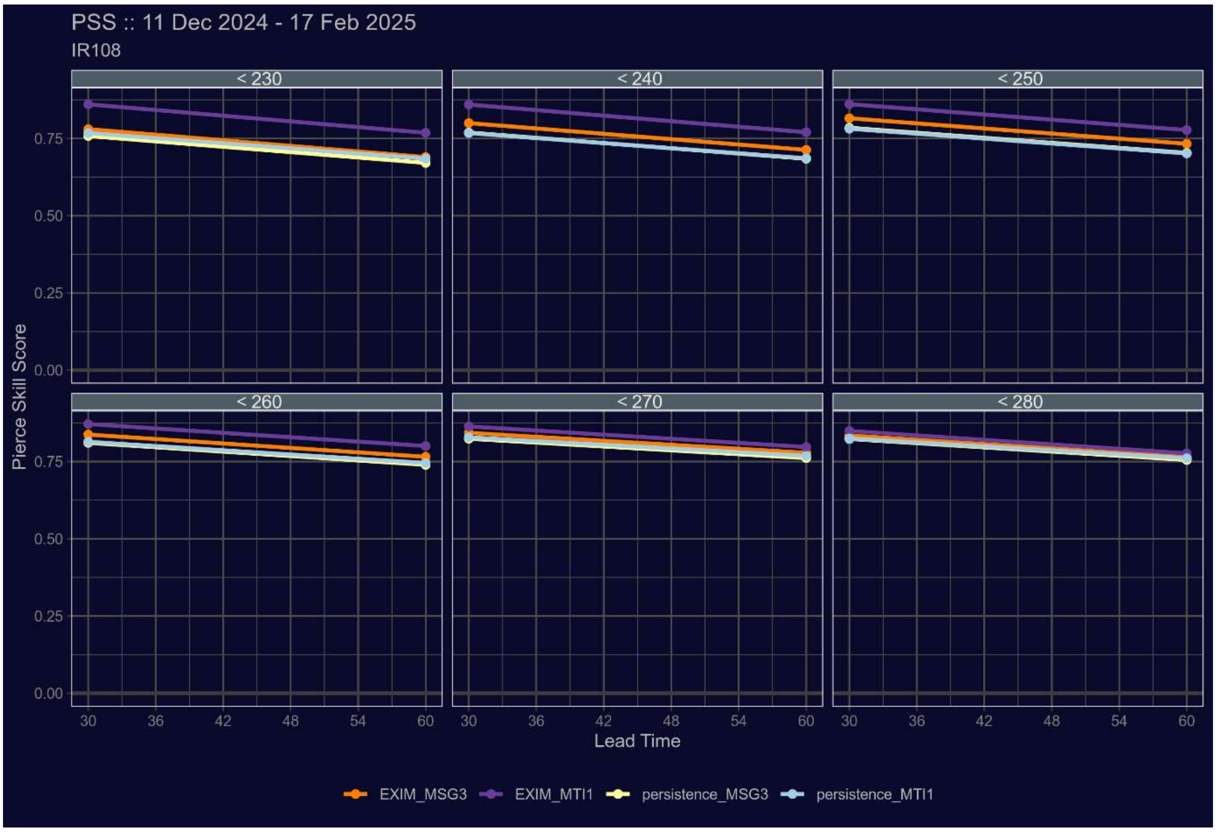

***Figure 3.20**: The PSS for IR108 is shown, with orange representing EXIM from MSG, yellow representing persistence from MSG, purple representing EXIM from MTG-I1, and light blue representing persistence from MTG-I1. The x-axis displays the lead times, with MTG-I1 and MSG plotted for direct comparison at their common lead times of 30 and 60 min. The panels show the results for the thresholds 230, 240, and 250 (top left to right), and 260, 270, and 280 (bottom left to right)*

3.4.2 CMA: Cloud Mask

The product cloud mask has two categories: cloudy (1) and not-cloudy (0). The threshold scores are calculated by looking for values equal to the categories.

Figure 3.21 shows that the FSS for *CMA* is very high, approaching 1. When considering the useful FSS, it becomes evident that part of this high value is due to chance, but not entirely. There is still a reasonable agreement between the two satellites, with values of 0.35 and 0.15 for category 0 and 1, respective. More importantly for this evaluation report, there is no difference between the lead times, indicating that the *EXIM* algorithm performs similarly for both satellites.

In **Figure 3.22** shows the PSS. For the MTG-I1 satellite, the PSS of the *EXIM* forecast is slightly improved compared to its persistence for lead time 30 minutes. For the MSG satellites, however, there is almost no difference between the forecast PSS and that of persistence. The improvement in *EXIM* forecasts with MTG-I data is attributed to a reduction of POFD (not shown). Additionally, the FAR is slightly reduced for both category 0 and category 1 at lead time 30 minutes.

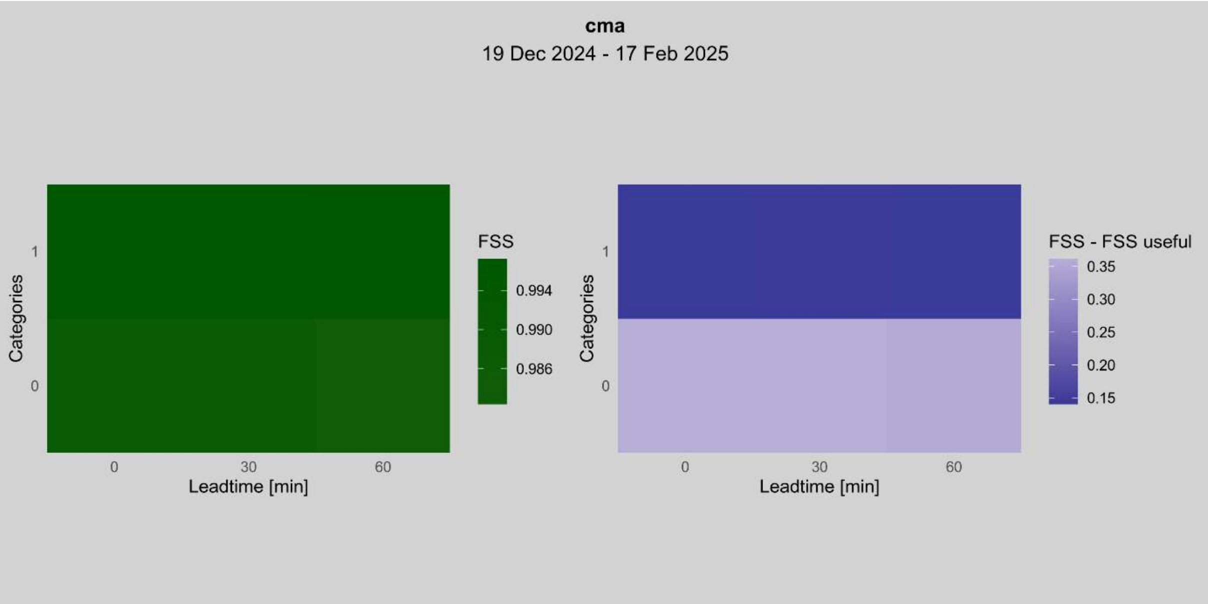

Figure 3.21 *Same as* ***Figure 3.19****, but for the cloud mask. The categories are 0 and 1.*

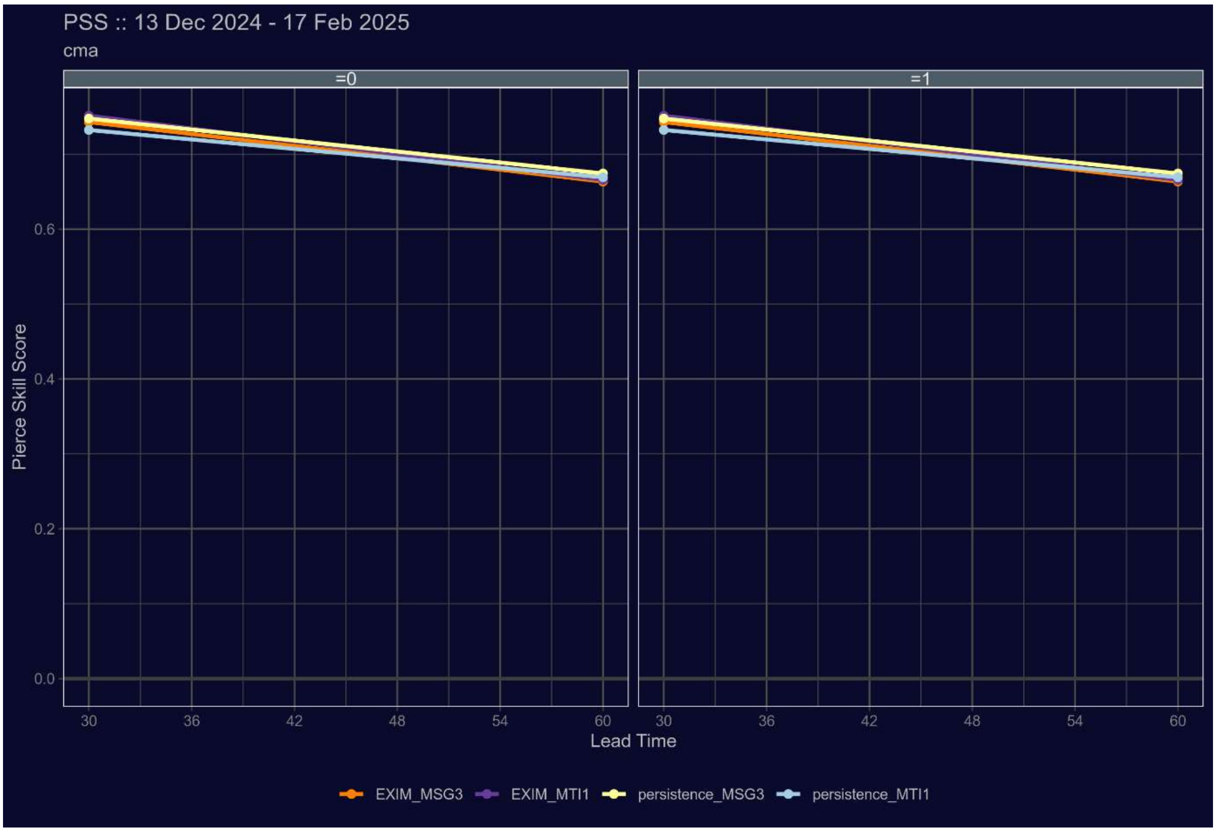

Figure 3.22*: Same as* ***Figure 3.20****, but for the cloud mask. The panels display the results for category 0 (left) and category 1 (right).*

3.4.3 CMIC phase: Cloud Microphysics - cloud phase

For NWCSAF product cloud microphysics (*CMIC phase*), the categories are liquid (1), ice (2), mixed (3), cloud-free (4) and un-defined (5).

FSS has high values (above 0.9) for the categories “liquid”, “ice”, “cloud-free” and “un-defined” as depicted in **Figure 3.23**. Only for the category “mixed”, the FSS shows a slightly lower agreement (0.65). When considering the useful FSS and removing the pure chance, one can see that the agreement is still quite good (0.35). As well for category 3, there is still some agreement as it is on the positive scale (0.1). FSS and $FSS - FSS_{useful}$ do not change with increasing lead times.

The improvement by *EXIM* forecasts compared to its persistence for satellite MTG-I1 can be seen in **Figure 3.24**. The degree of improvement differs with the categories. While there is some improvement for categories 2, 3 and 5, the improvement is small for category 1. For all categories, the improvement is rather visible for MTG-I than for MSG. The improvement of PSS is driven by a reduction of POFD (not shown). FAR is reduced in the same pattern as PSS is increased.

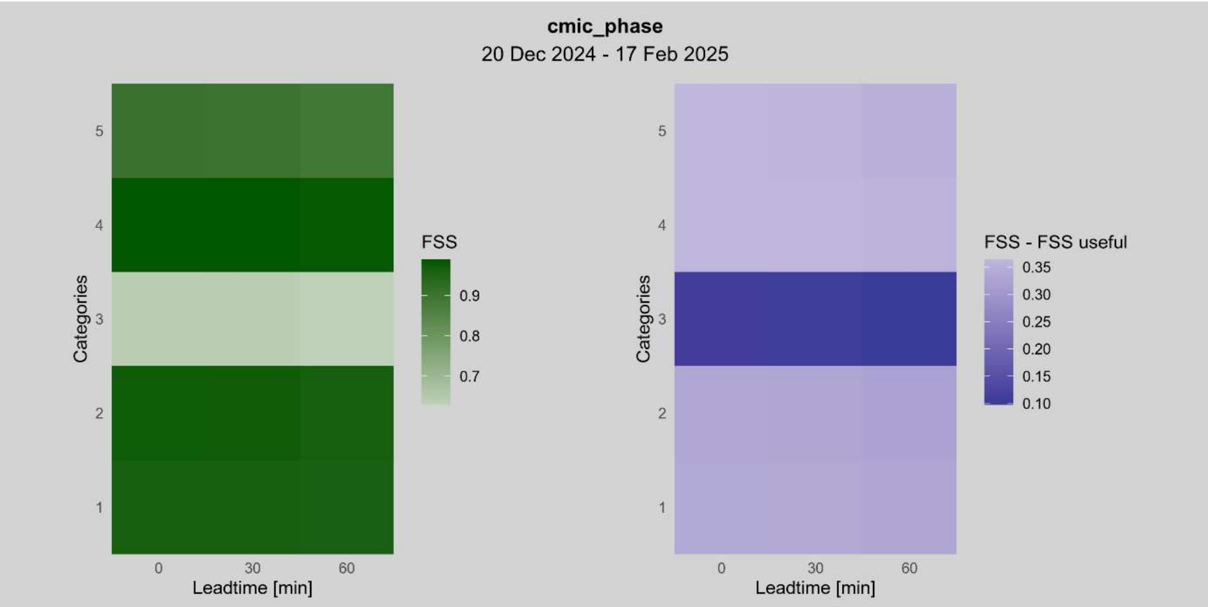

***Figure 3.23**: Same as **Figure 3.19**, but for the CMIC phase. The categories are liquid (1), ice (2), mixed (3), cloud-free (4), and un-defined (5).*

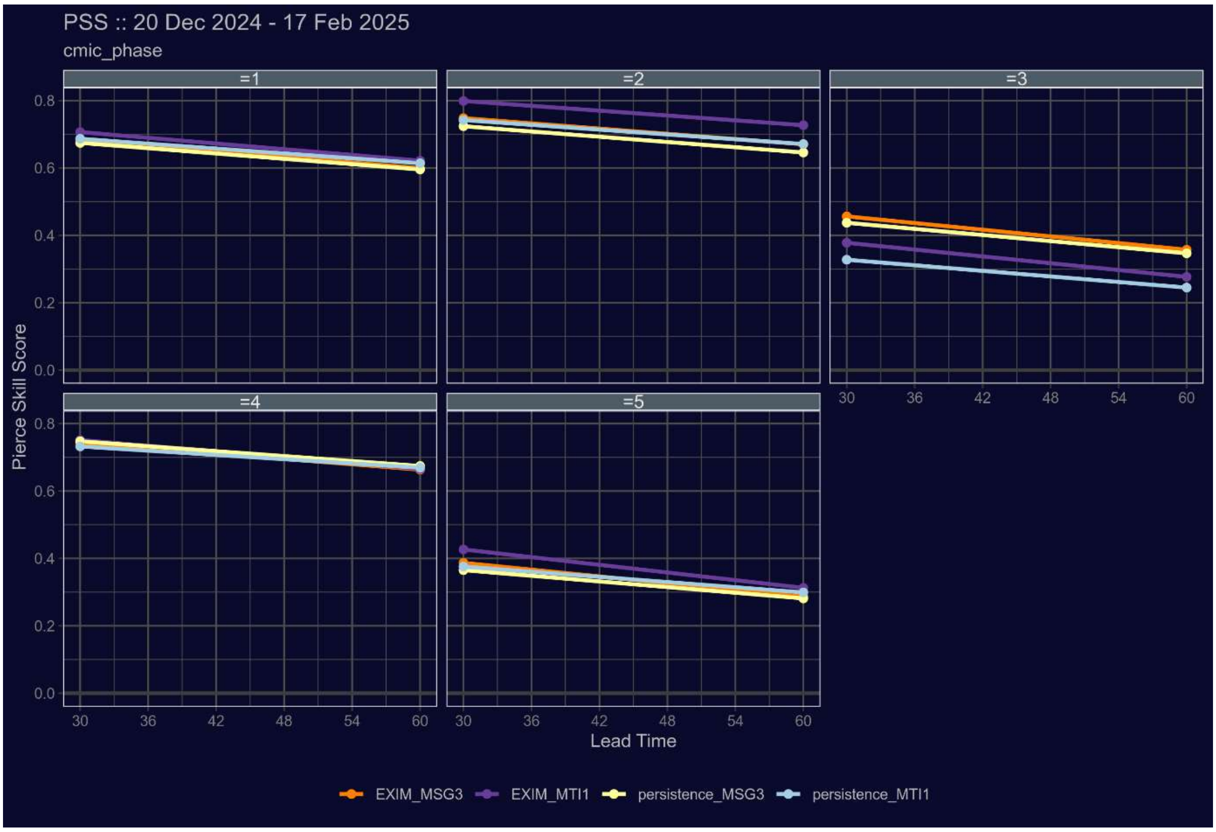

***Figure 3.24**: Same as **Figure 3.20**, but for the CMIC phase. The panels display the results for the categories: liquid (1), ice (2), mixed (3) (top from left to right); cloud-free (4), and un-defined (5) (bottom from left to right).*

3.4.4 CTTH alti: Cloud Top Temperature and Height - altitude

The cloud top height (*CTTH alti*) product is verified for thresholds ranging from greater than 1,000 m to 11,000 m in 1,000 m increments.

Figure 3.25 displays the FSS for *CTTH alti*, showing values close to 1 for most thresholds. Only at high altitudes (thresholds of 10,000 m and 11,000 m) do the values slightly decrease, with lowest FSS around 0.8. After removing the randomness factor by subtracting FSS_{useful}, a reduction in agreement between the two satellites becomes apparent. This suggests that a small portion of the high FSS values result from random chance, but a notable agreement remains, with values ranging between 0.2 and 0.4.

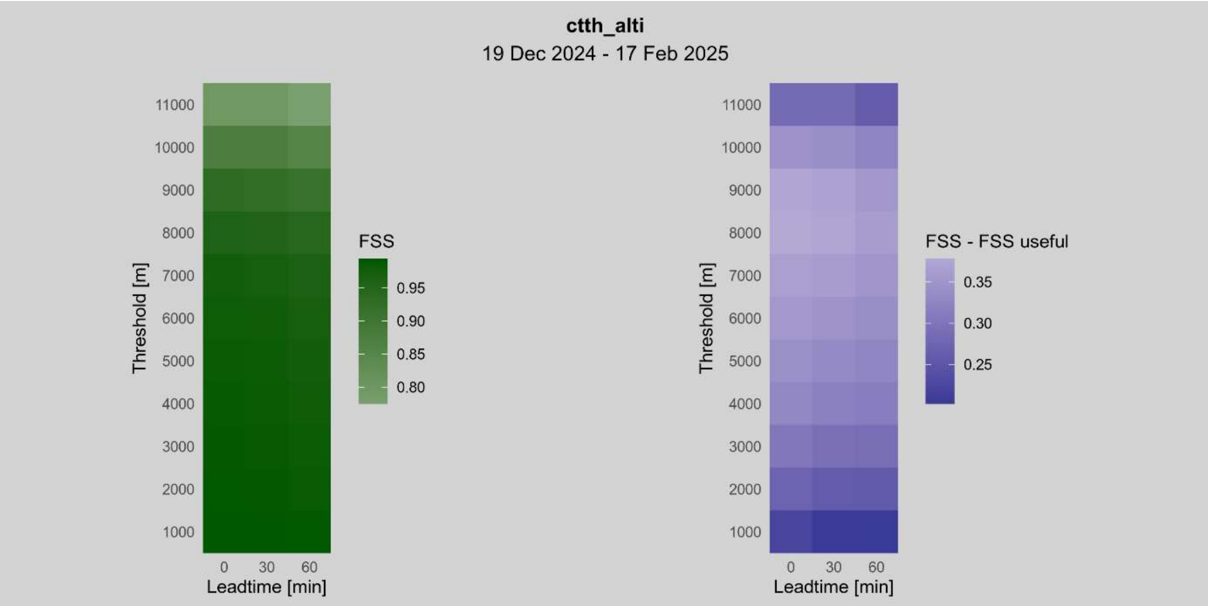

***Figure 3.25**: Same as **Figure 3.19**, but for the CTTH alti. The thresholds range from 1,000 m to 11,000 m in 1,000 m steps.*

When comparing the *EXIM* forecasts for the satellites with their respective persistence at lead time 30 and 60 minutes, the behaviour for the two satellites is similar. RMSE is lower in *EXIM* forecasts compared to their persistence (see **Figure 3.26**), with a more pronounced reduction for MTG-I1.

While there is not much differences visible for the score PSS (not shown), FAR (see **Figure 3.27**) is reduced (improved) by *EXIM* forecasts compared to their persistence for thresholds greater than 6000 m. The difference is more clearly visible for MTG-I1 than for MSG.

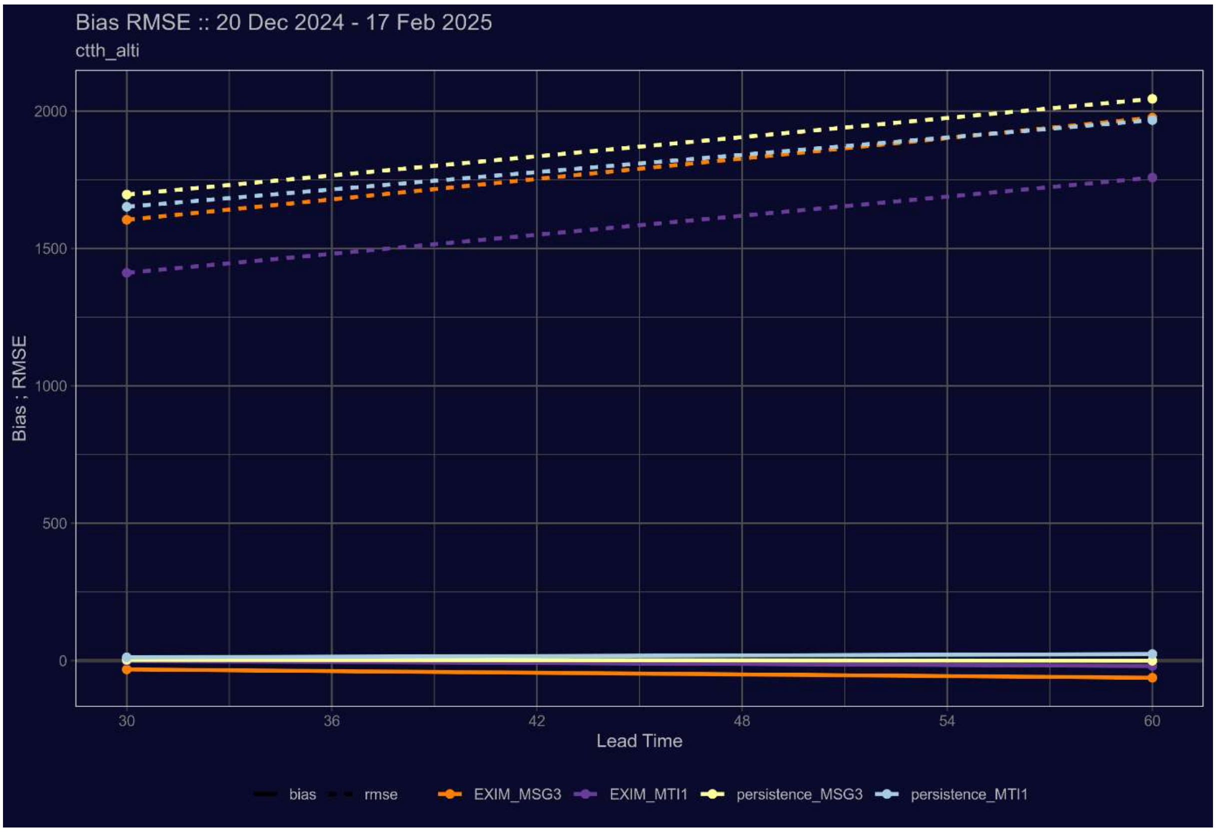

***Figure 3.26**: Same as **Figure 3.20**, but for the CTTH alti and for RMSE and BIAS. The panel display scores for the thresholds starting at 1,000 m (top left) to 11,000 m (bottom right) in 1,000 m steps.*

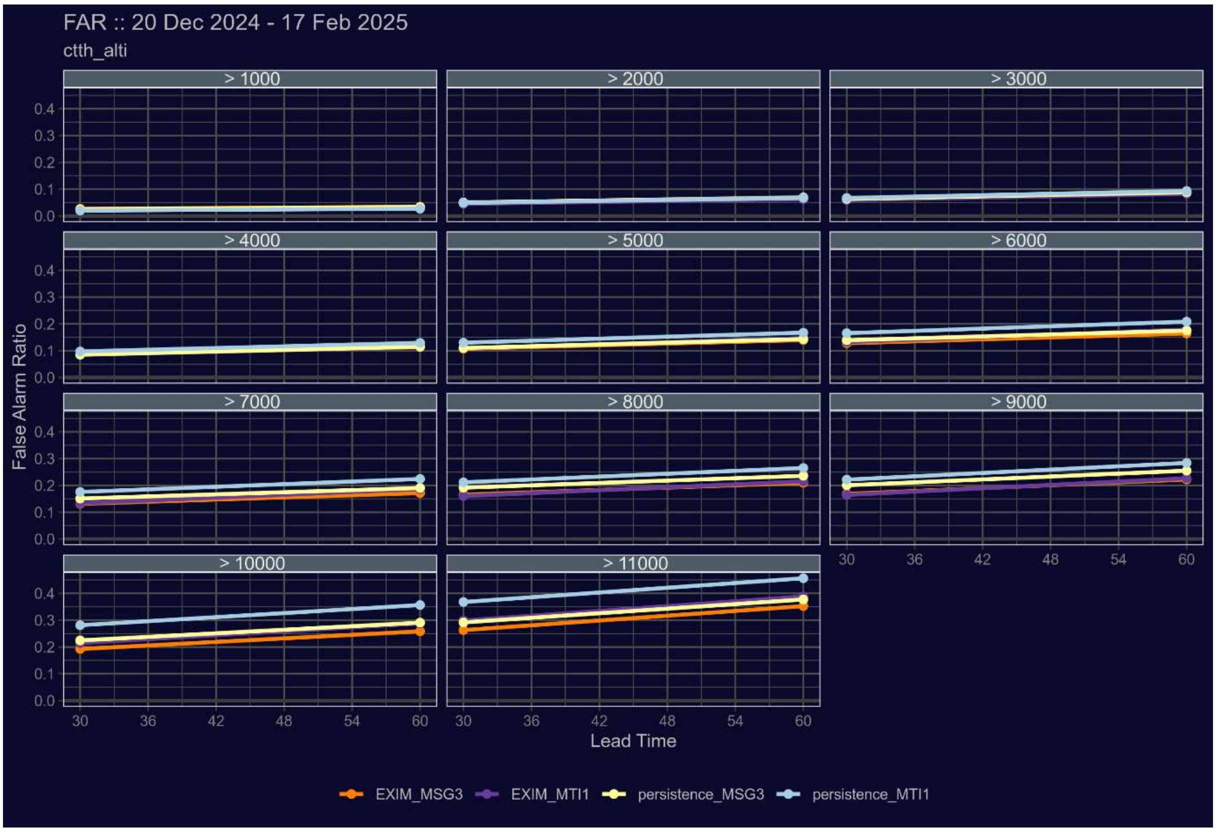

***Figure 3.27**: Similar as **Figure 3.20**, but for the CMIC phase and for the FAR score. The panels display scores for the thresholds starting at 1,000 m (top left) to 11,000 m (bottom right) in 1,000 m steps.*

3.4.5 CT: Cloud Tpye

The scores for product cloud type (CT) are plotted for the following categories: cloud-free land (1), cloud-free sea (2), snow over land (3), sea ice (4), very low clouds (5), low clouds (6), mid-level clouds (7), high opaque clouds (8), very high opaque clouds (9), fractional clouds (10), high semi-transparent thin clouds (11), high semi-transparent moderately thick clouds (12), high semi-transparent thick clouds (13), high semi-transparent above low or medium clouds (14), high semi-transparent above snow ice (15). The surface types are excluded for the grid-point-based evaluation. The threshold scores were calculated by looking for values equal to the specified categories.

Figure 3.28 shows the agreement between MTG-I1 and MSG *EXIM* forecasts for *CT*, as derived by FSS. Most categories show good FSS values close to 1. However, there are a few exceptions. Category 11 has a lower FSS value (0.7) compared to the other categories. Another notable exception is category 15, where the complete failure is due to no cases being observed in the MSG forecasts. In the case of MTG-I1, only very few cases were detected. As a result, this category has too few cases to be properly evaluated.

When examining the point-based scores comparing *EXIM* forecasts of the two satellites with their respective persistence, it is evident that the PSS (see **Figure 3.29**) improves in *EXIM* forecasts compared to persistence for both satellites, with a maximum improvement of up to 0.15 for category 9 and MTG-I1. This improvement is primarily driven by a better POD, though there is also a small improvement in POFD (not shown). As seen with other scores and products, the improvement in *EXIM* forecasts compared to persistence is more significant for MTG-I1 than for MSG. The only exception is category 5, where no difference is observed. Additionally, the FAR (not shown) is reduced in *EXIM* forecasts compared to persistence for all categories except 5 and 6, with the reduction being more pronounced for MTG-I1.

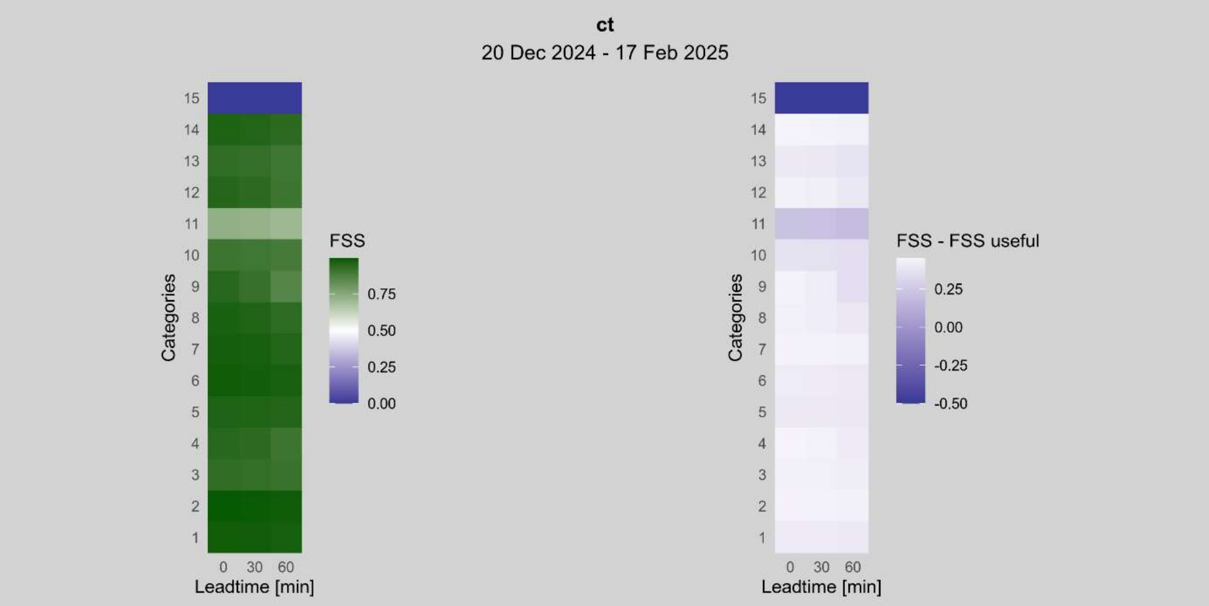

***Figure 3.28**: Same as **Figure 3.19**, but for the CT. The displayed categories are: cloud-free land (1), cloud-free sea (2), snow over land (3), sea ice (4), very low clouds (5), low clouds (6), mid-level clouds (7), high opaque clouds (8), very high opaque clouds (9), fractional clouds (10), high semi-transparent thin clouds (11), high semi-transparent moderately thick clouds (12), high semi-transparent thick clouds (13), high semi-transparent above low or medium clouds (14), high semi-transparent above snow ice (15).*

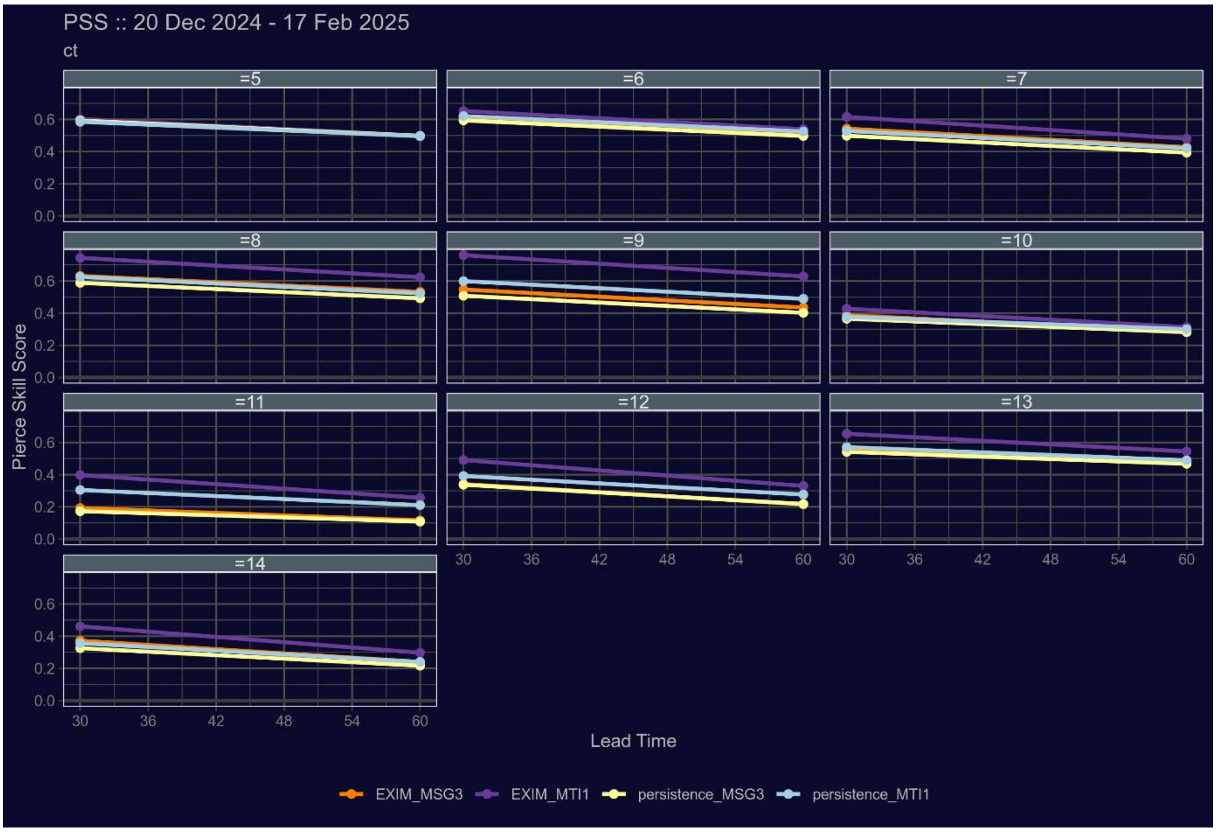

***Figure 3.29**: Same as **Figure 3.20**, but for the CT. The panels display scores for the categories: very low clouds (5), low clouds (6), mid-level clouds (7), high opaque clouds (8), very high opaque clouds (9), fractional clouds (10), high semi-transparent thin clouds (11), high semi-transparent moderately thick clouds (12), high semi-transparent thick clouds (13), high semi-transparent above low or medium clouds (14). Categories (1), (2), (3) and (4) were not included as they are surface types. Category (15) is left out due to an insufficient number of cases for comparison.*

3.4.6 CRRPh: Convective Rainfall Rate based on Cloud Physical properties

The forecasts of the product "Convective rainfall rate based on cloud physical properties – intensity" (*CRRPh*) are evaluated using the following thresholds 0.2, 2, 3, 4, 5, 6, 7, 8 and 9 mm/h.

EXIM forecasts for both satellites show an improvement in RMSE compared to their respective persistence (see **Figure 3.30**). The PSS (not shown) for *EXIM* forecasts using the MTG-I1 satellite improves by about 0.5 at a lead time of 30 minutes compared to persistence. This improvement is primarily due to a better POD (not shown). For MSG, *EXIM* forecasts show a similar improvement in POD compared to persistence, but only for the smallest threshold. The FAR (see **Figure 3.31**) of *EXIM* forecasts is lower for both satellites compared to their respective persistence.

The standard FSS is not too helpful in a situation where a systematic difference in the numerical values of the compared data sets is present. As we observed this for CRRPh(MSG) vs. CRRPh(MTG), and also for the PCPh product discussed in the next section, this metric is not discussed for the "physical" precipitation products.

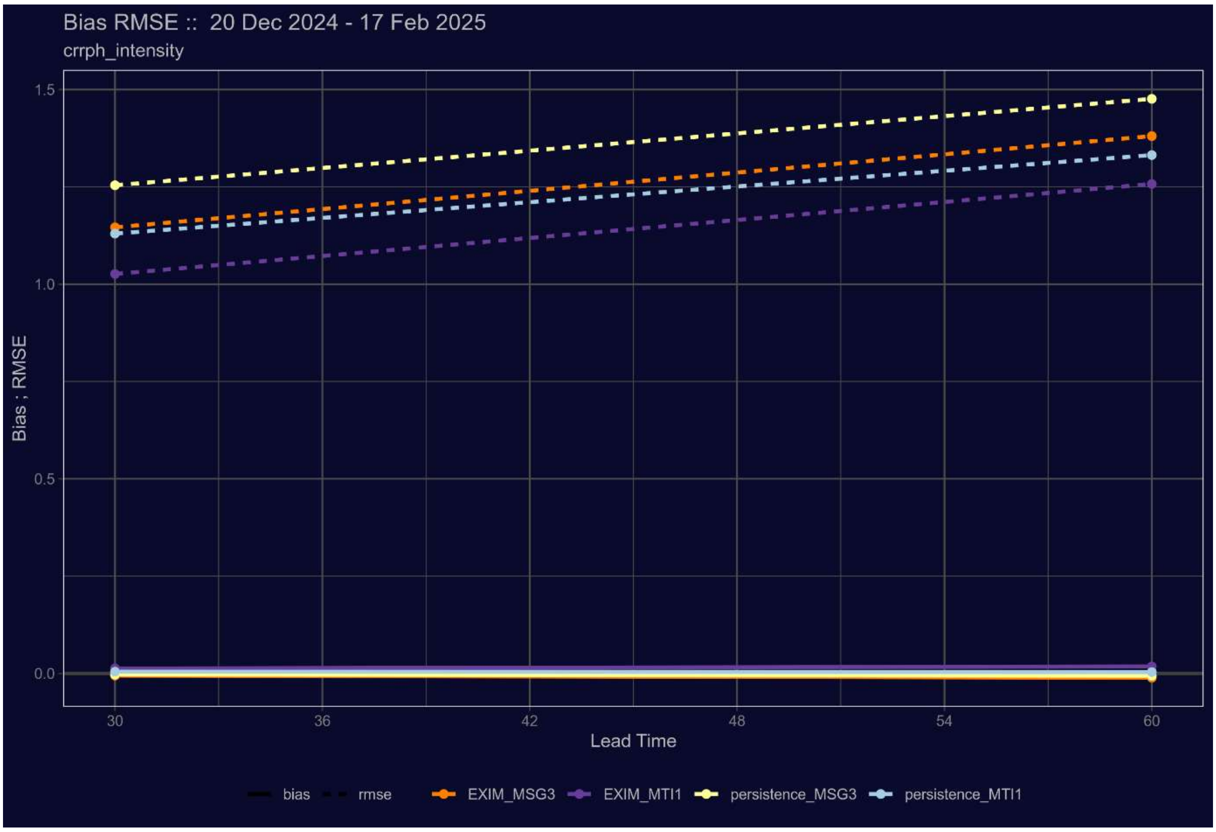

Figure 3.30*: Same as* ***Figure 3.20****, but for the CRRPh intensity and the scores Bias (solid lines) and RMSE (dashed lines). The panel display scores for 0.2, 2, 3, 4, 5, 6, 7, 8 and 9 mm/h (top left to bottom right).*

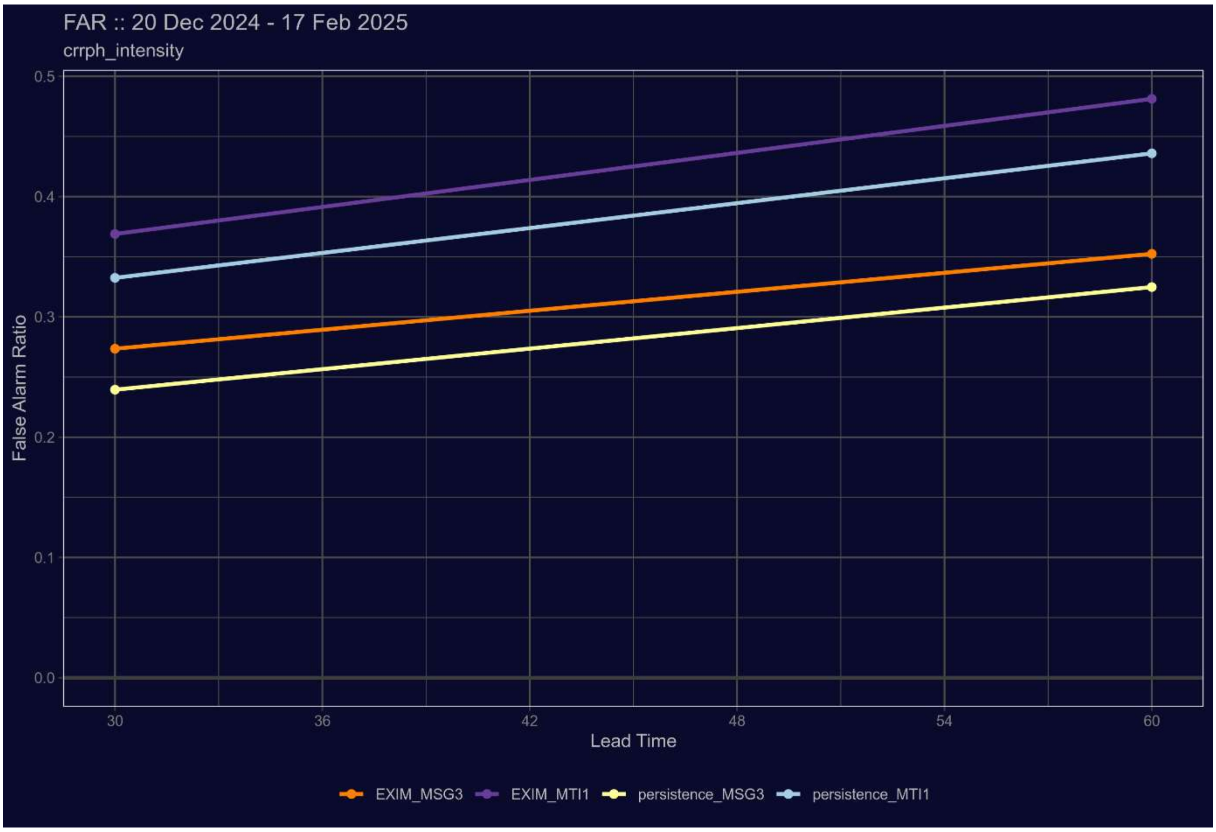

***Figure 3.31**: Same as **Figure 3.20**, but for the CRRPh intensity and the score FAR. The panel display scores for 0.2, 2, 3, 4, 5, 6, 7, 8 and 9 mm/h (top left to bottom right).*

3.4.7 PCPh: Precipitating Clouds based on Cloud Physical Properties

The product "Precipitating Clouds Based on Cloud Physical Properties" (*PCPh*) is evaluated for precipitation probabilities of 1 %, 11 %, 21 %, 31 %, 41 %, 51 %, 61 %, and 71 % based on threshold scores. For these threshold scores, values exceeding the specified thresholds were analysed.

When comparing the common lead times 30 and 60 minutes, RMSE (see **Figure 3.32**) is for both satellites likewise slightly reduced by about 1 for *EXIM* forecasts compared to their respective persistence. In this evaluation, the improvement of PSS (not shown) by *EXIM* forecasts compared to its persistence is mainly visible for MTG-I1. The increase of PSS is driven mainly due to an improvement of POD. There is a small improvement of POFD (not shown) visible for MSG by *EXIM* forecasts compared to its persistence which is not visible for MTG-I1. While the smallest two chosen thresholds (1 % and 11 % precipitating probability) are slightly higher or equal to for higher thresholds onwards (21 % up to 71 %), FAR (Figure 3.33) starts to be improved by *EXIM* forecasts compared to their respective persistence for both satellites. The greater the thresholds the greater the improvement.

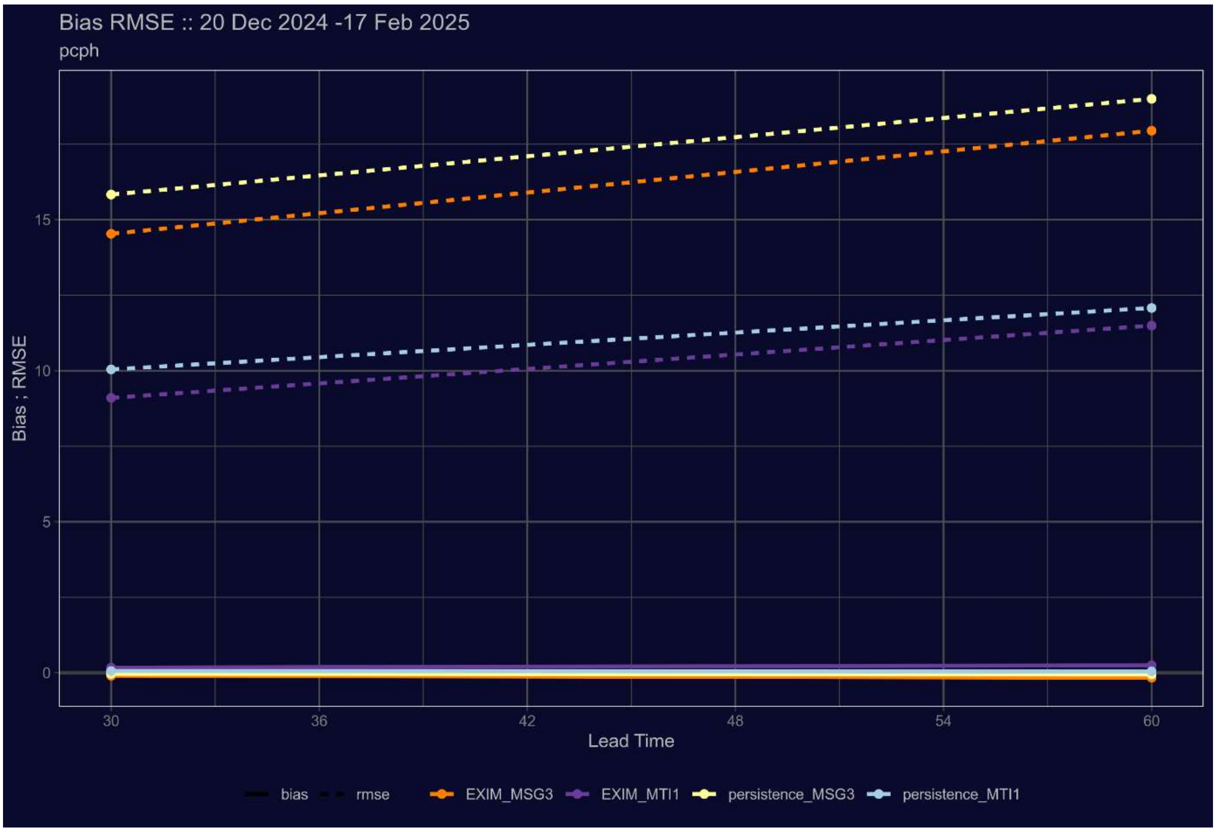

***Figure 3.32**: Same as **Figure 3.20**, but for the PCPh for the score RMSE (dashed line) and BIAS (solid line). The panel display scores for the thresholds 1, 11, 21, 31, 41, 51, 61, 71 % (top left to bottom right).*

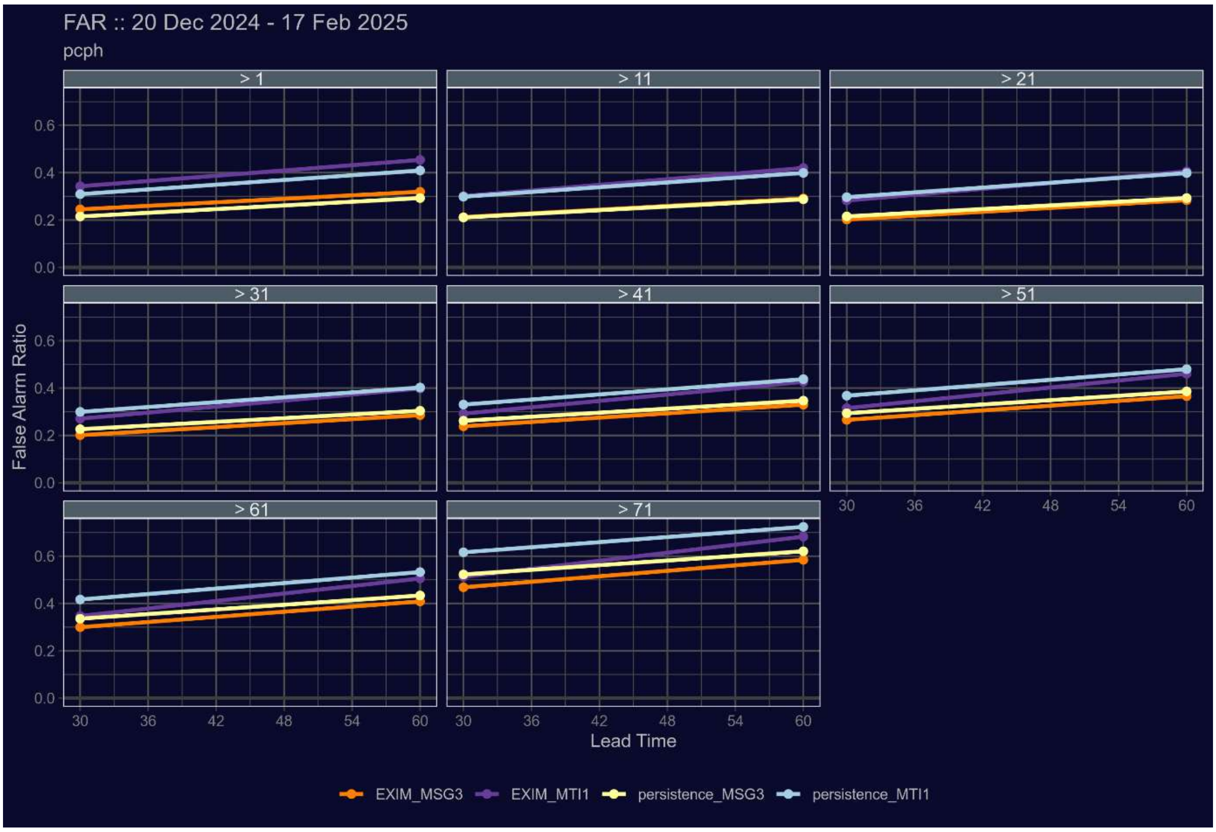

Figure 3.33*: Same as* ***Figure 3.20****, but for the PCPh and for score FAR. The panel display scores for the thresholds 1, 11, 21, 31, 41, 51, 61, 71 % (top left to bottom right).*

4. CONCLUSIONS

This validation covers four main aspects. First, whether the two modified products and the newly added one can be beneficially extrapolated by *EXIM*, and how *EXIM* performs for the MTG-I1 satellite compared to *EXIM* forecasts for MSG. Second, how the CTTH filter performs compared to previous extrapolation setups. Third, whether water vapour atmospheric motion vectors shall be added to the list of used AMVs. And fourth, how EXIM forecasts using satellite data from MTG-I1 compare to data from MSG. As a reminder, only the chapter “comparison of MTG-I1 against MSG” compares both satellites, while the other three use data from MSG only. Summary conclusions based on the detailed results shown in Chapter 3 are presented hereafter.

4.1 EXIM COMPARISON AGAINST PERSISTENCE

The forecasts of *EXIM* are validated with each change in the algorithm of an input product or when a new product is added to the set of extrapolated products in *EXIM*.

The underlying results confirm that forecasts of *EXIM* add value for five of the investigated products (*HRVIS, VIS06, VIS08, CRRPh and CMIC cot*) compared to persistence, and they all reach the threshold accuracy of being “*on average better than persistence forecast*”.

Only *ASII-TF* does not improve the scores being forecasted with *EXIM* compared to persistence. Tropopause foldings are typically located where there is vertical shearing at the jet stream combined with the ageostrophic convergence of polar, subtropical, and stratospheric air masses. Being driven by other mechanisms than the horizontal wind fields, it is hardly surprising that *ASII-TF* cannot be reasonably forecasted with *EXIM*.

4.2 CTTH FILTER AND ITS RECOMMENDATION

The re-evaluation of the CTTH filter for the products *CTTH alti*, *IR108* and *IR38* with an expansion of the boundaries and therefore including the lowest levels just above the surface does not yield different results than the previous evaluation. For those height-dependent products, “control” continues being the best option and “2-layer, with filter” being the second best option.

However, for *CT* and *CMIC phase* this matter changes when surface types, and type without assigned height (*CT*) and cloud-free and un-defined categories (*CMIC phase*) are not considered in the multi-category evaluation. Depending on whether a complete image is wanted or whether only the clouds types/ phases are of interest either “control” or “2-layer, with filter” is recommended.

4.3 WV VECTORS NOT RECOMMENDED

The conclusion on whether to add WV vectors to the set of AMVs used for the extrapolation field for *CMIC cot* is the same as it has been for all of the other products. It is not recommended adding the AMVs from those two channels since they do not improve the forecasts.

4.4 COMPARISON OF MTG-I1 VS. MSG

To summarize the results of the comparison of MTG-I1 and MSG, we conclude that the algorithm of *EXIM* can be applied in the same way. The scores are agreeing mostly very well compared between the two satellites. Comparing the forecasts for both satellites to their respective persistence shows the same behaviour or even an improvement for MTG-I1 compared to MSG.

5. OUTLOOK

This underlying validation displayed that forecasts of *EXIM* are outperforming persistence for the implemented products and satellite channels for all lead times. However, the dominance is decreasing with increasing lead time. For all moving pixels, which should be the majority, *EXIM* will get better with time. Only stationary events and at the boundaries of the domain where the movement origins from, persistence will dominate with increasing lead time. The reason why the improvement by forecasts of *EXIM* is getting smaller with increasing lead time is not fully understood. The investigation of this question is foreseen to be addressed in future validations.

While the extrapolation of *CMIC cot* has been included already in NWC/GEO v2025, the flexibility with respect to image resolution (thus enabling HRVIS extrapolation), the cosine correction of VIS channels and the option of user-defined lead times were postponed to the subsequent release because of superordinate project considerations.

6. ANNEX

6.1 ROC CURVE

The relative operating characteristic (ROC) curve (**Figure 6.1**) depicts the relation of POD and POFD. Therefore, PSS can be inferred, with PSS being 1 in the top left corner and decreasing vertically to the dashed line. At the dashed line PSS = 0 and the forecast has no skill.

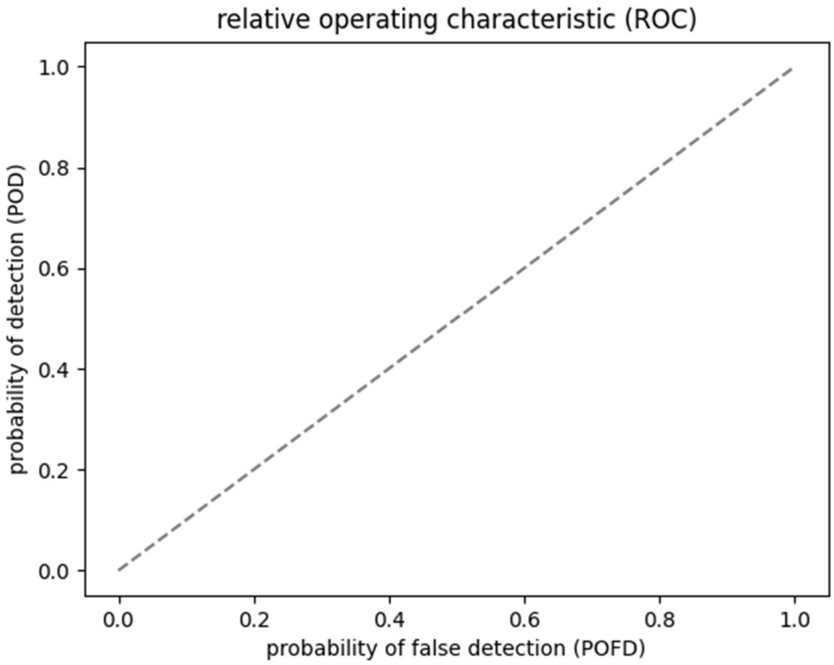

Figure 6.1: *Exemplary ROC curve without data.*

6.2 PERFORMANCE DIAGRAM

The performance diagram (**Figure 6.2**) summarises the scores POD, FAR, CSI and BIAS in one plot. A perfect forecast would be in the top right corner; the worst scenario would be in the bottom left corner.

POD		Probability of detection. Increasing from bottom to top.
FAR		False Alarm Ratio. Increasing from right to left.
BIAS	- - -	Bias. The diagonal is bias-free (1.0). Lower values express too few (forecasted) “yes”-events and higher values too many (forecasted) “yes”-events.
CSI	___	Critical Success Index. Curved lines. Increasing from bottom left corner to top right corner.

For a more comprehensive description, please refer to: https://www.cawcr.gov.au/projects/verification/Roebber/PerformanceDiagram.html

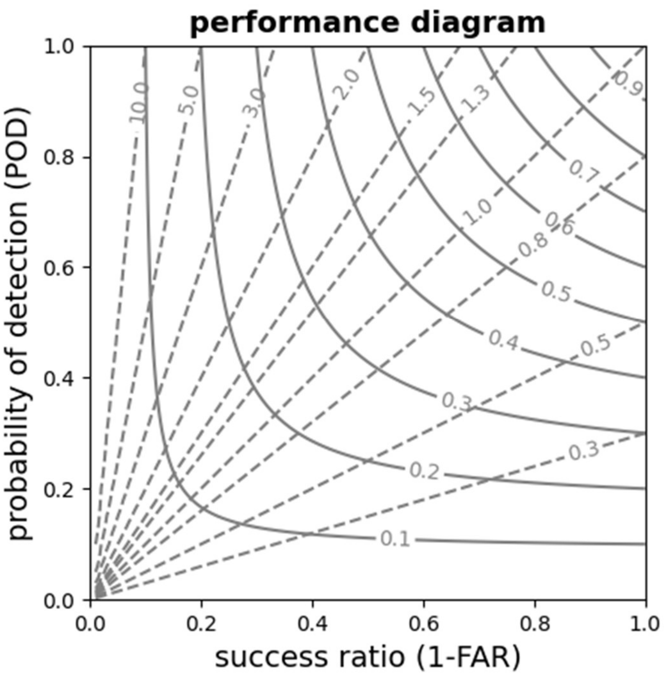

Figure 6.2: *Exemplary performance diagram without data.*