Comparing AIFS and high-resolution IFS (4.4km) simulations for 2-m temperature extremes: when does the ML model excel and when does resolution matter?

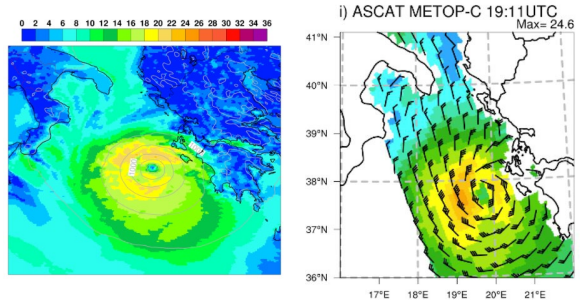
Estíbaliz Gascón

Evaluation Section

Extreme weather evaluation: Complementary strategies

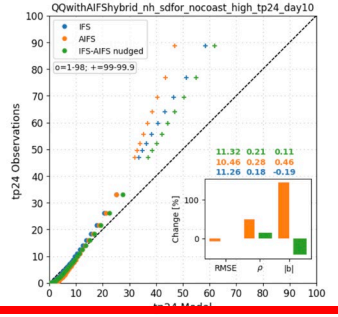
Evaluation of extreme case studies

- Evaluation of past case studies
- Continuous evaluation of emerging events (Daily Report, Severe event catalogue)



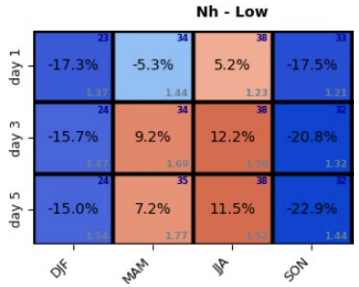
Verification of the climatology of extremes

- To evaluate how much the model climatology fits the observations climatology
- i.e. Quantile-quantile comparison focused on extremes (highlighting >99th percentile)



Verification with specific scores for extremes

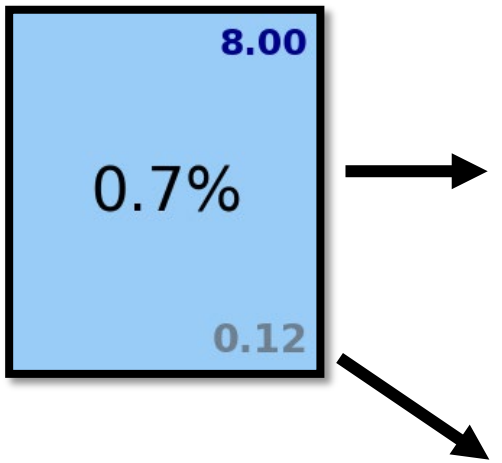
- We must choose useful verification scores for extremes
- How to define extremes? Percentiles or fixed thresholds
- Evaluation against point observations (SYNOP/HDOBS)



SCORECARDS for EXTREMES TOOL

GOAL: compare the model performance for extremes for surface variables between 2 models

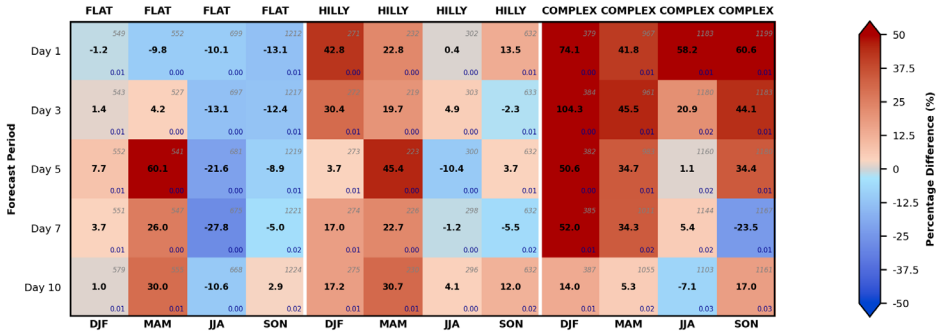
BOLD BLACK FRAME: the differences are statistically significant



Number of cases

% of improvement or degradation (model 2 compared to model 1)

Score value of model 2



SCORES:

- Threshold weighted mean absolute error (**twMAE)**
- Threshold weighted mean square error (**twRMSE**)
- Equitable Threat Score (**ETS**)
- Pierce Skill Score (**PPT**)

DEFINITION OF EXTREME:

- Percentiles from OBS clim (20 years)
- Fixed thresholds (i.e. 50 mm in 24h)

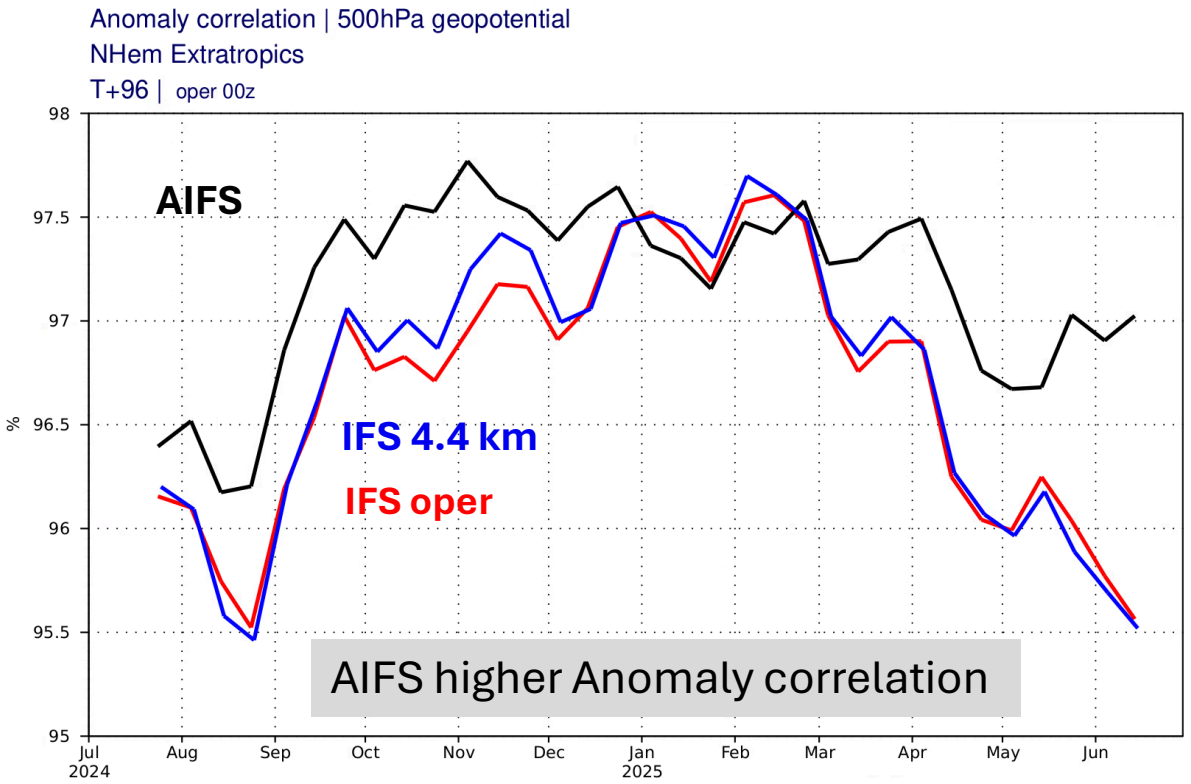
FURTHER DIAGNOSTIC TOOLS:

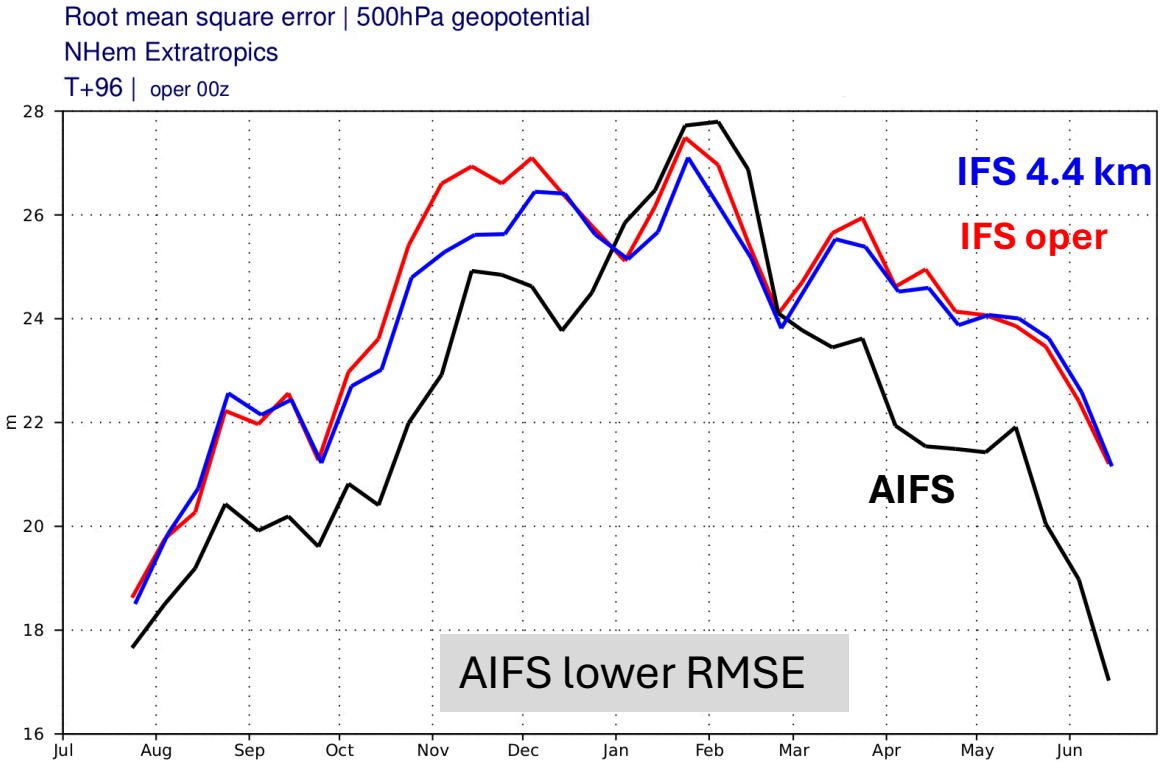
- Comparing different percentiles
- Frequency bias
- False alarms/misses/hits

Synoptic-scale verification

500 hPa geopotential height

- Slightly better performance of DestinE compared to IFS oper
- Much better performance of AIFS compared to the other two physical models





Why might these models predict extremes better than the operational IFS?

AIFS

- **Longer predictability** and less forecast jumpiness: fewer location errors. Higher scores in synoptic variables.

- **Trained with ERA5**: detection of observed extremes that are not easy to capture with the current IFS configuration and it corrects biases.

IFS 4.4 km

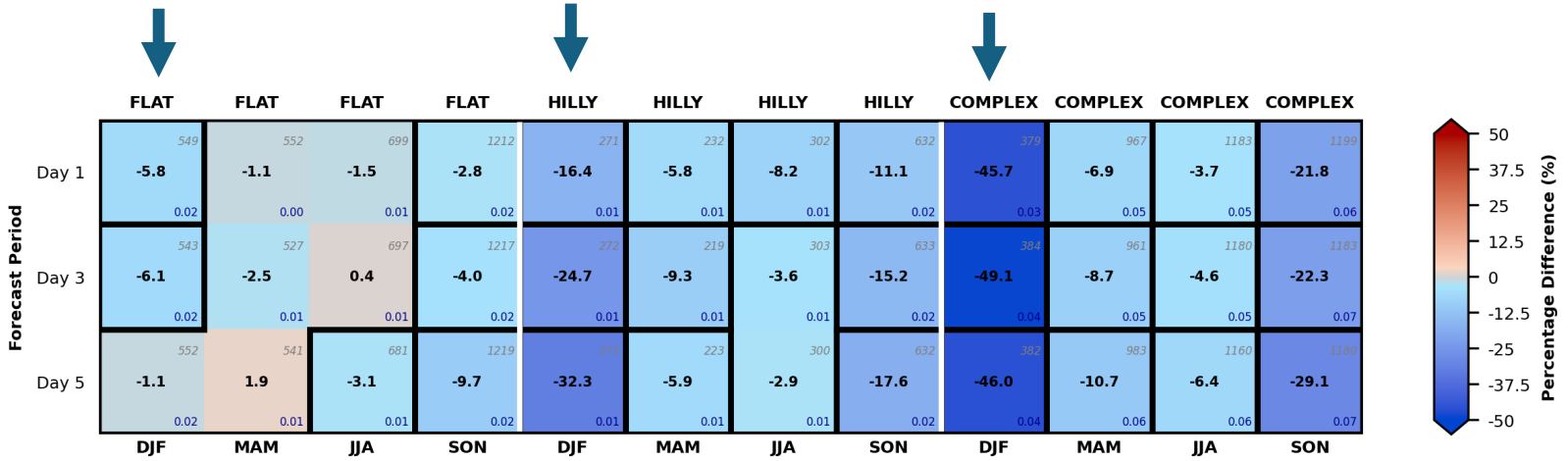
- **Higher horizontal resolution** helps to improve the prediction of extremes which depend on *small-scale processes* and orography

- **Physical processes** are generally better resolved with higher horizontal resolution forecasts: better prediction of extremes

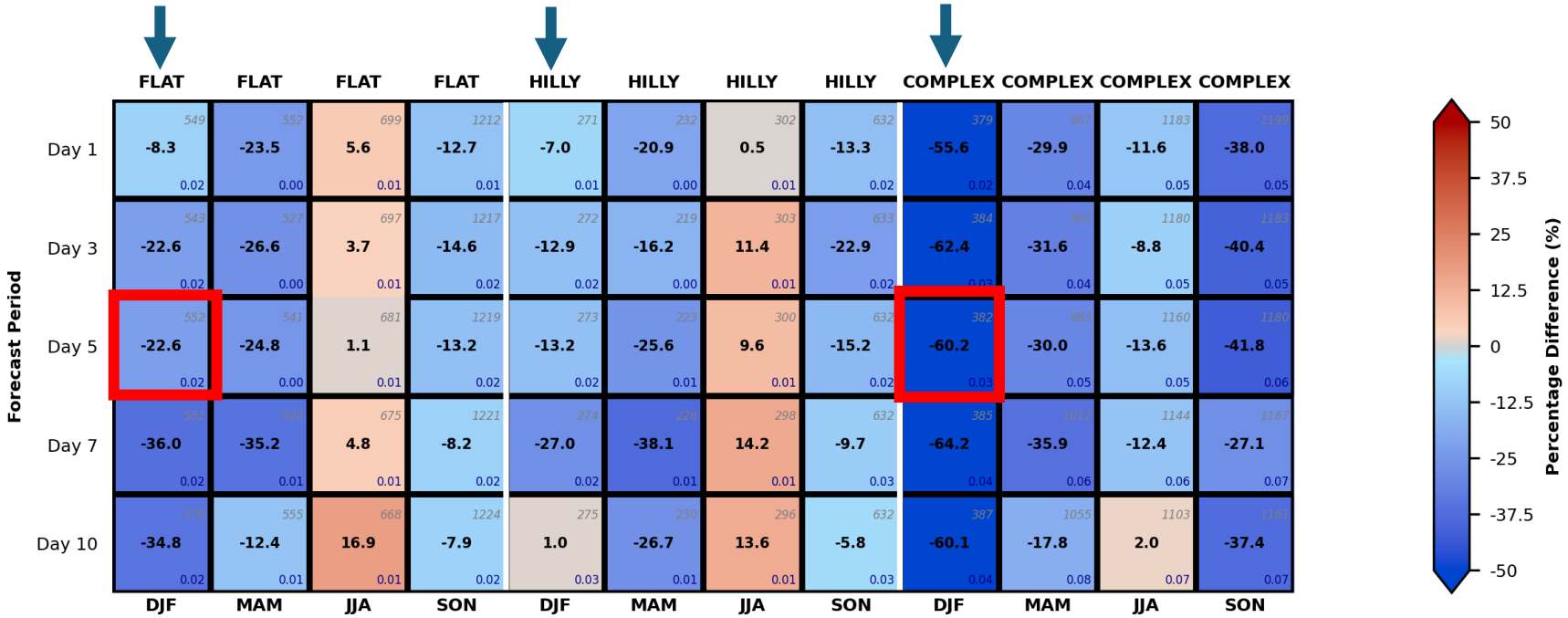
2m daily mean temperature cold extremes (1st percentile station-based)

twMAE in Europe

IFS 4.4 km vs. **IFS**
(ref)



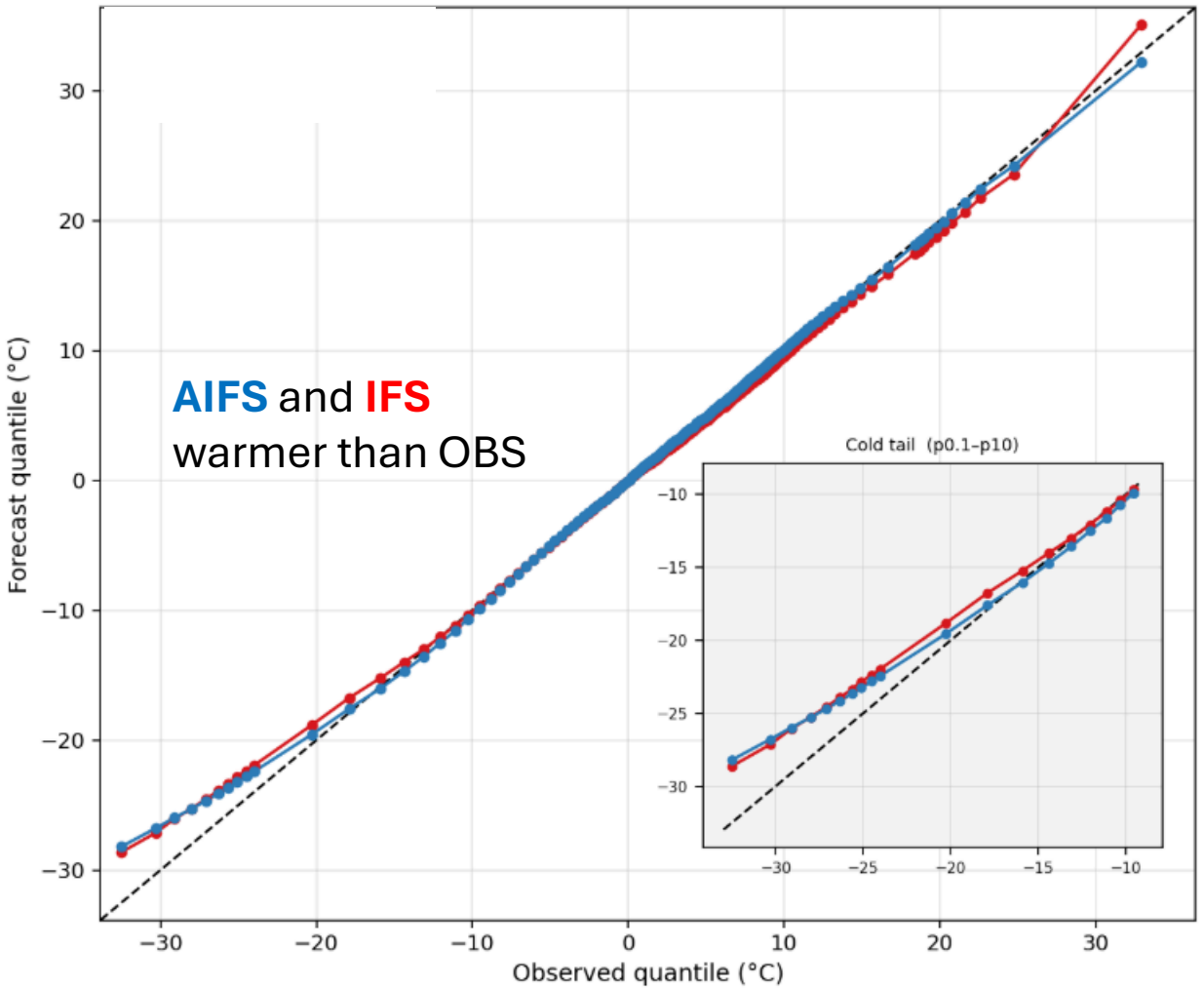
AIFS vs. **IFS**
(ref)



winter day 5

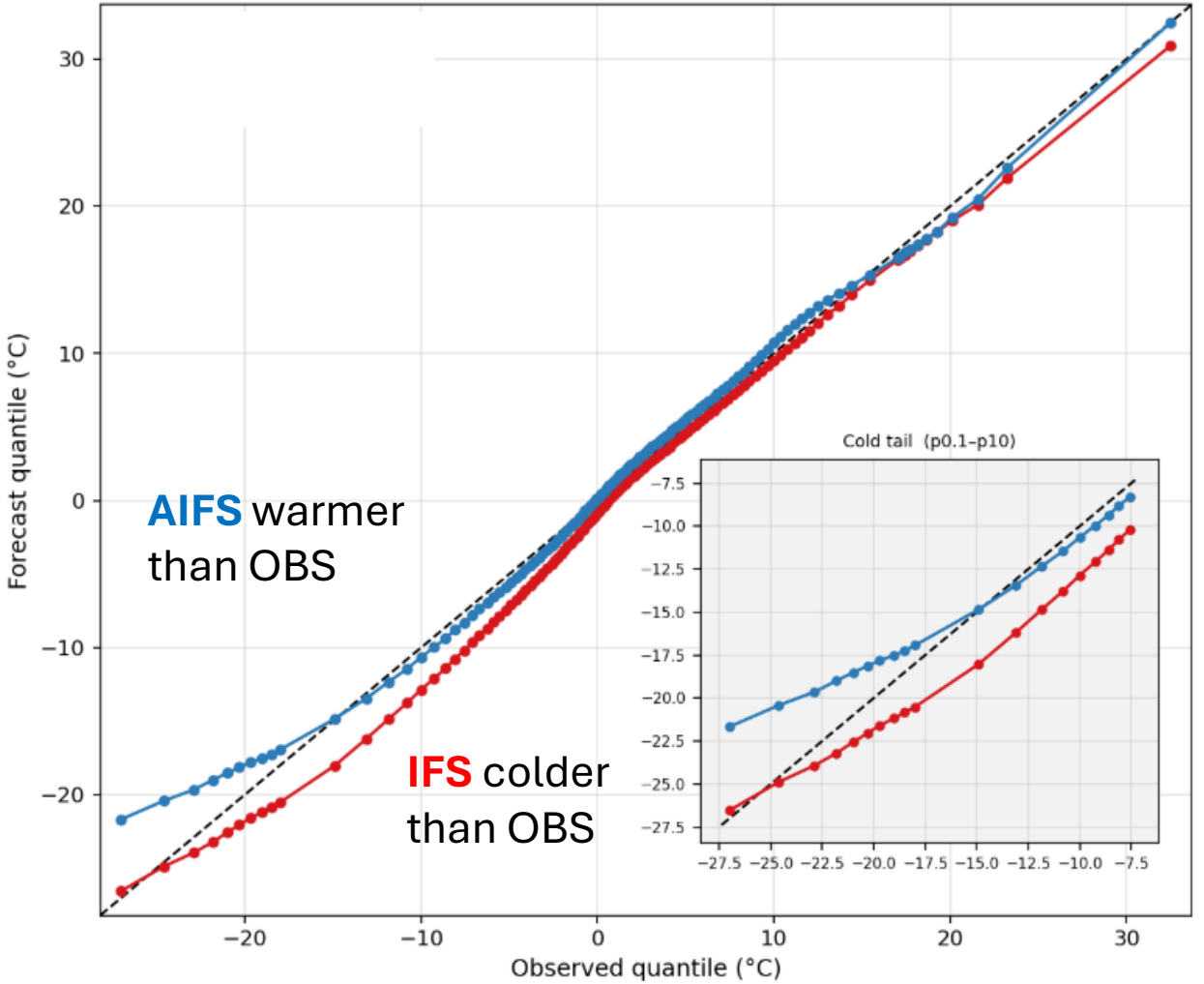
Flat areas

Are IFS errors larger under very stable conditions?



Complex orographic areas

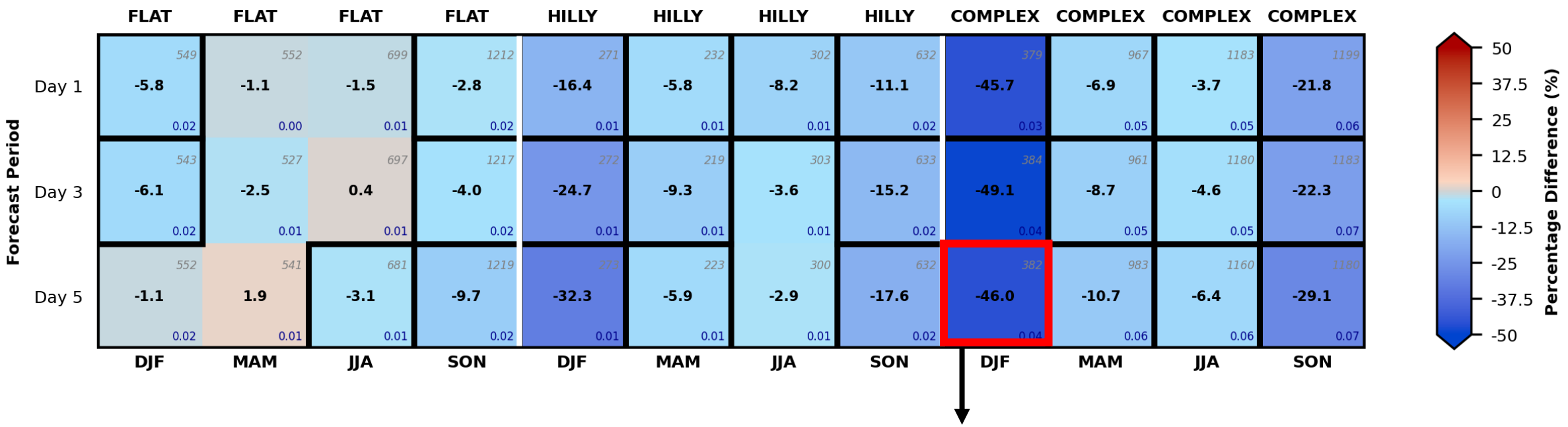
Is there a general cold bias in the IFS over the mountains?



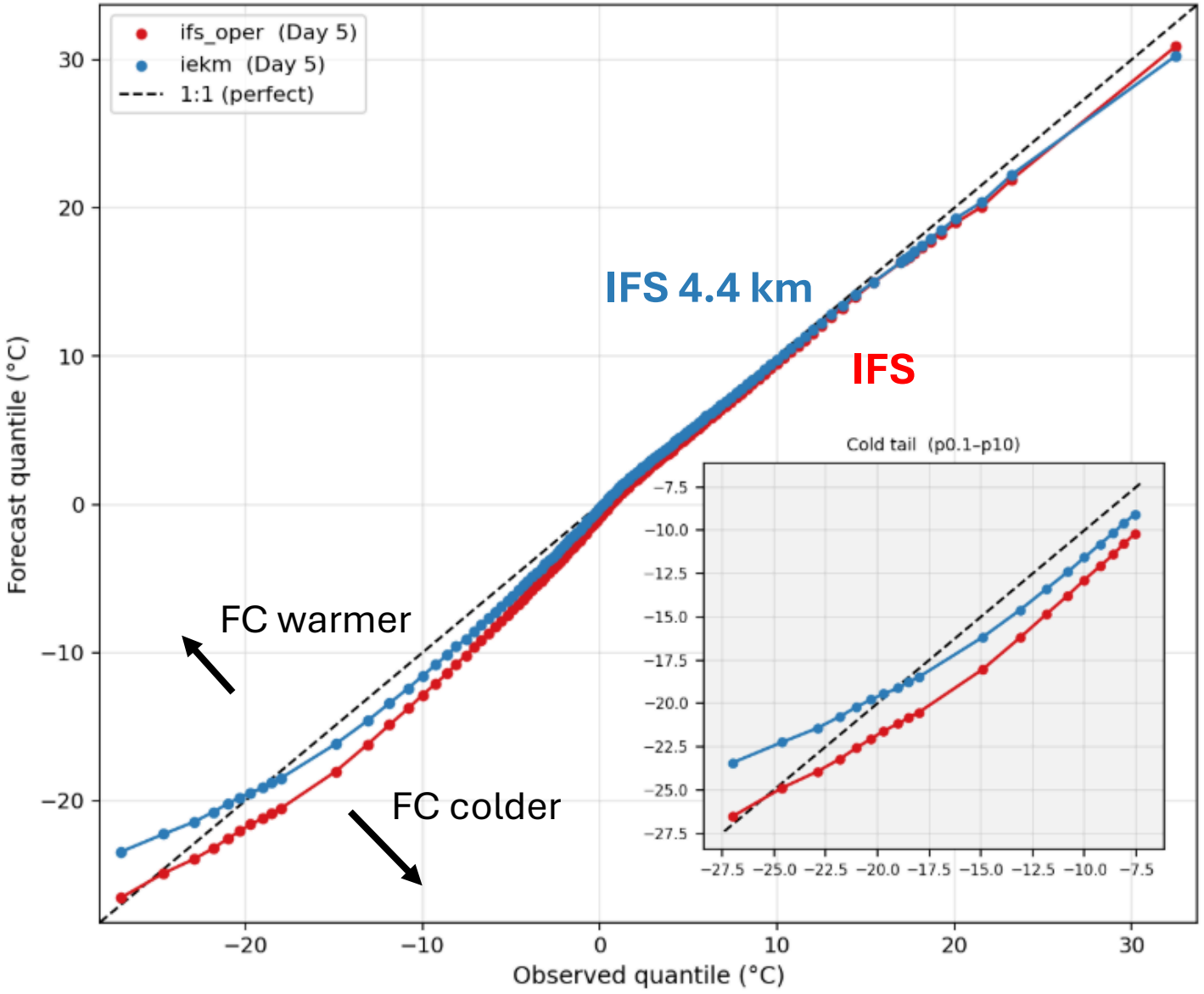
2 m temperature cold extremes

IFS 4.4 km vs. IFS

twMAE



IFS 4.4 km large improvements in mountainous areas



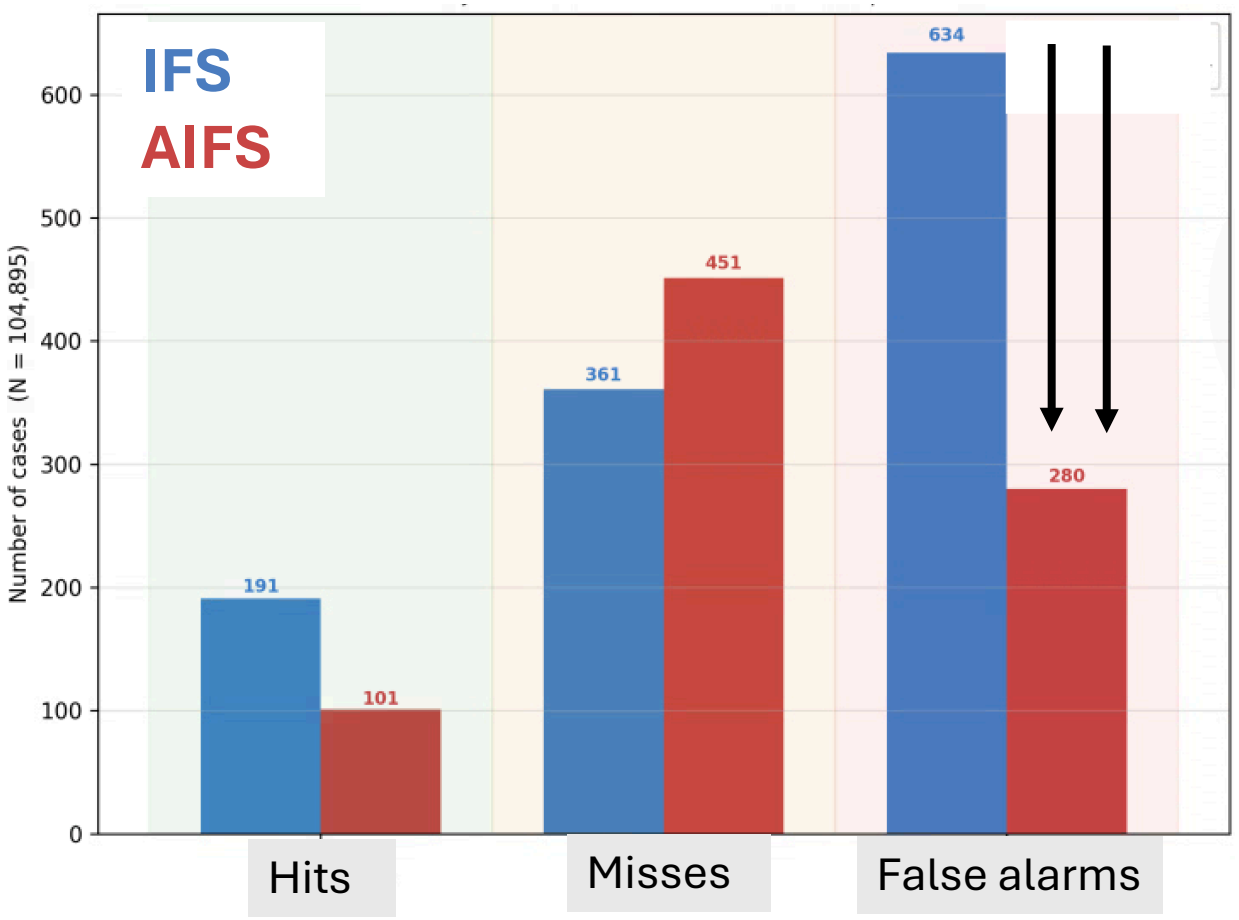
Complex orographic areas

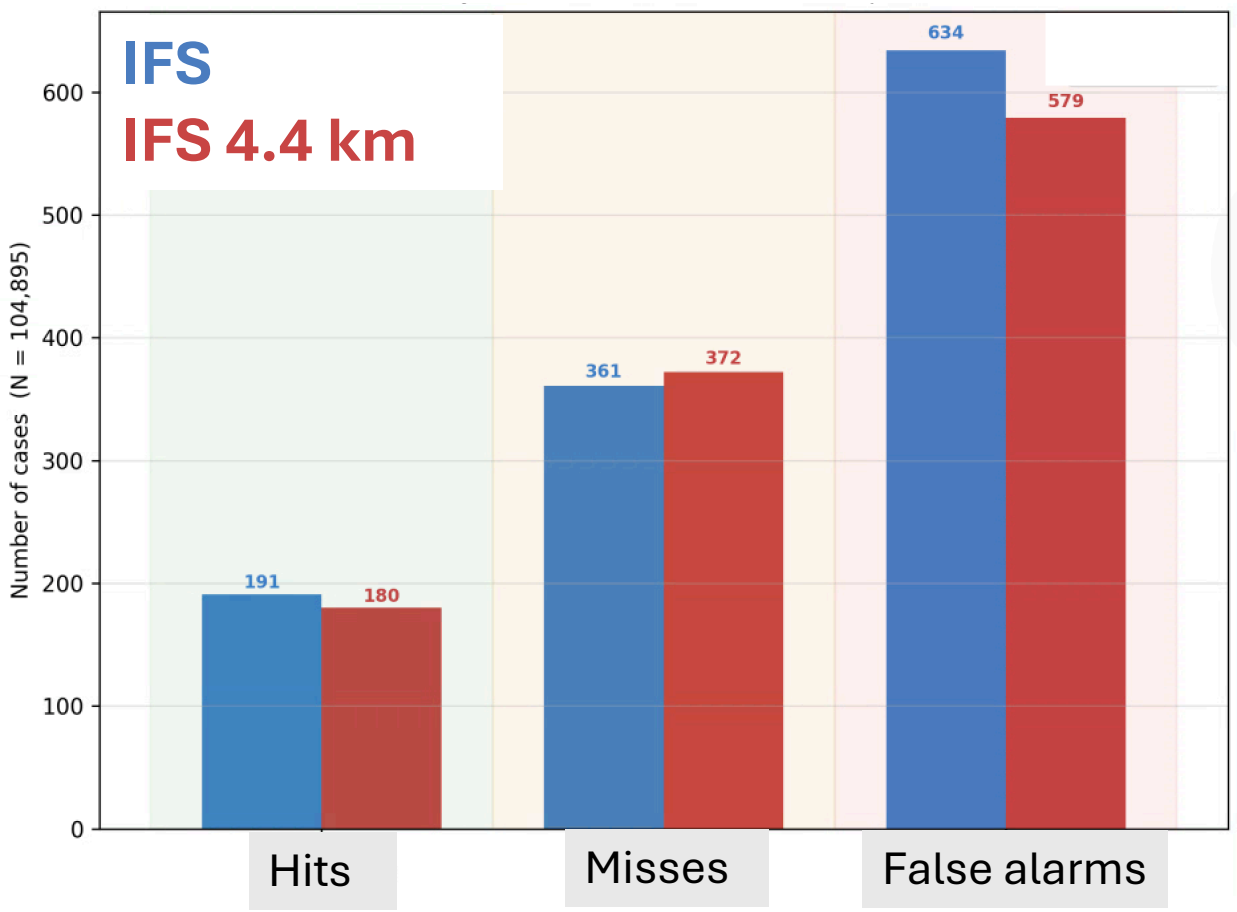
winter day 5

IFS 4.4 km slightly warmer than **IFS**: reduction of cold bias when increasing the resolution?

NOTE: IFS and IFS 4.4 km look pretty similar for flat areas

Why do we have much better forecasts with AIFS in flat areas?

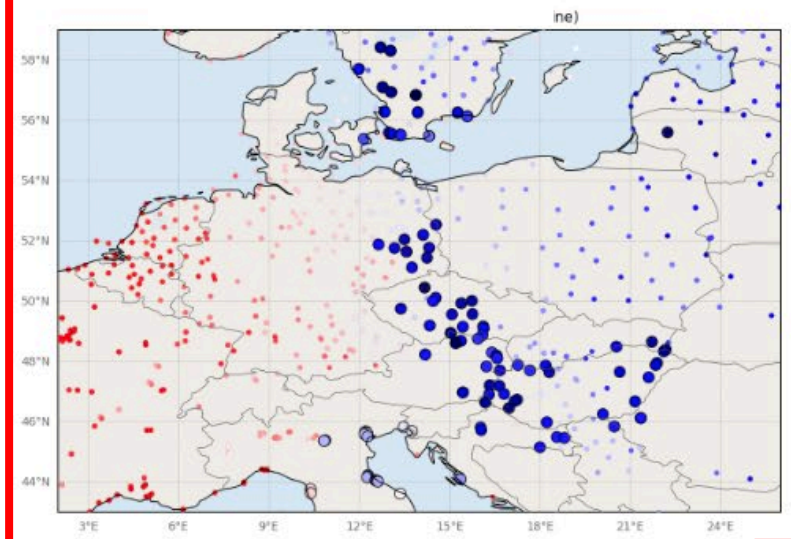




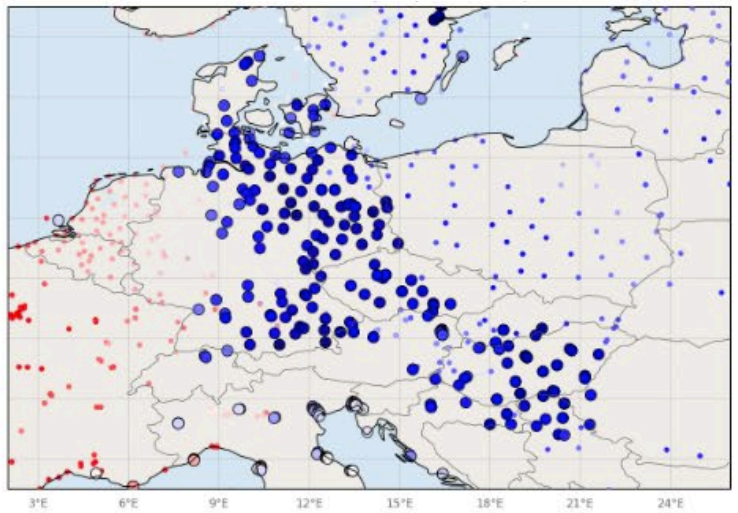
AIFS considerably **reduces false alarms** relative to IFS

Case study central Europe 08-01-2026, lead time day 5. *2m mean temperature*

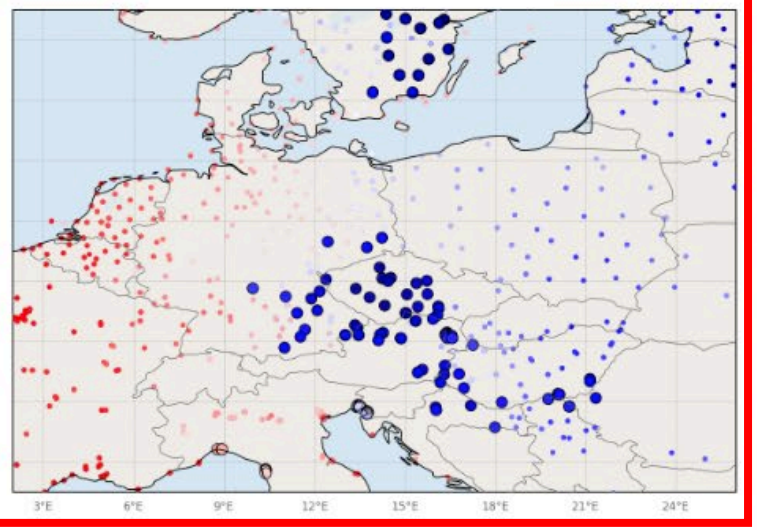
OBS

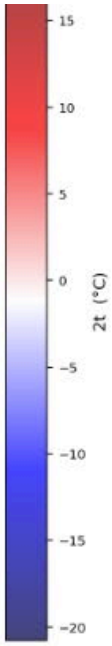


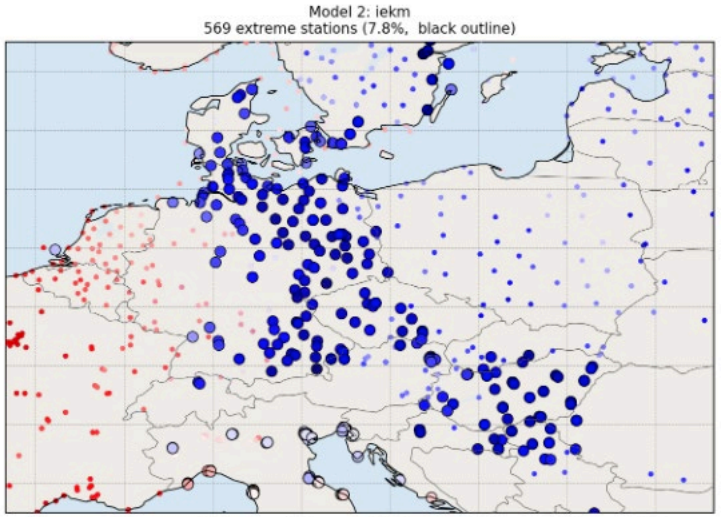
IFS



AIFS







More false alarms in the **IFS** and **IFS 4.4 km** than the AIFS

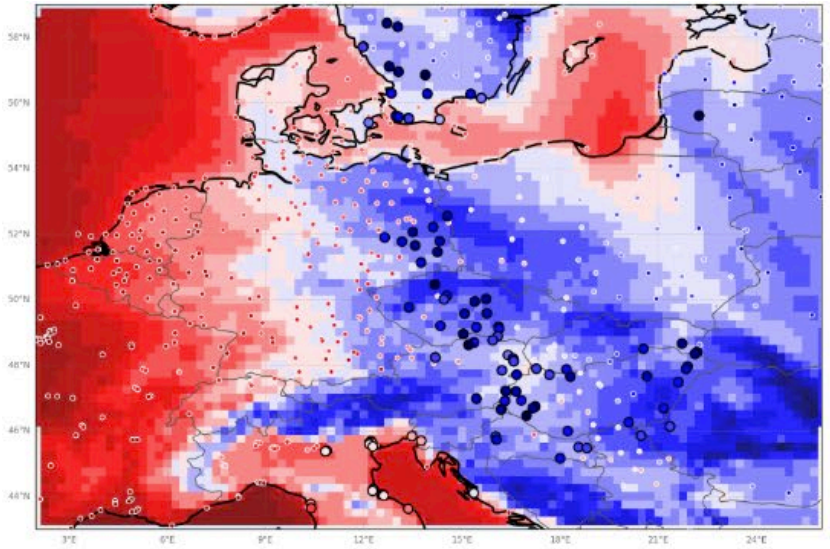
BIG BLUE CIRCLES: values below the 1st percentile of the obs climatology per station

IFS 4.4 km

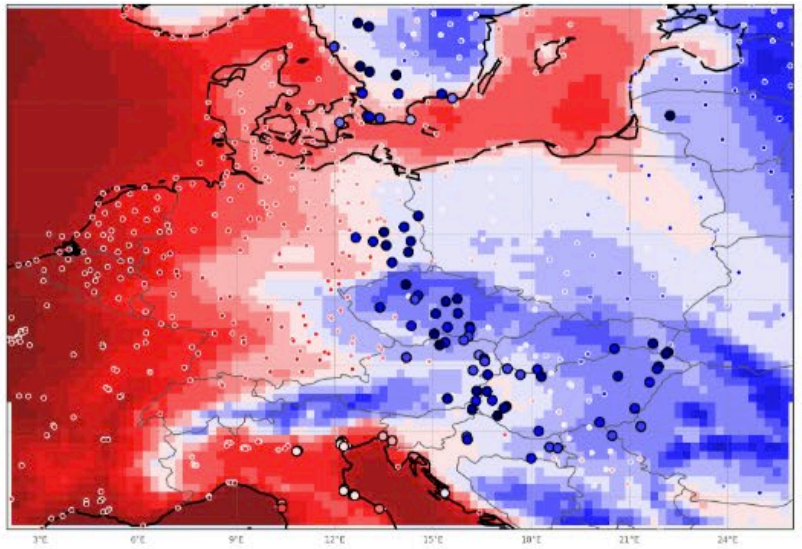
Case study central Europe 08-01-2026
2m mean temperature

BIG BLUE CIRCLES: OBS values below the 1st percentile of the obs climatology per station

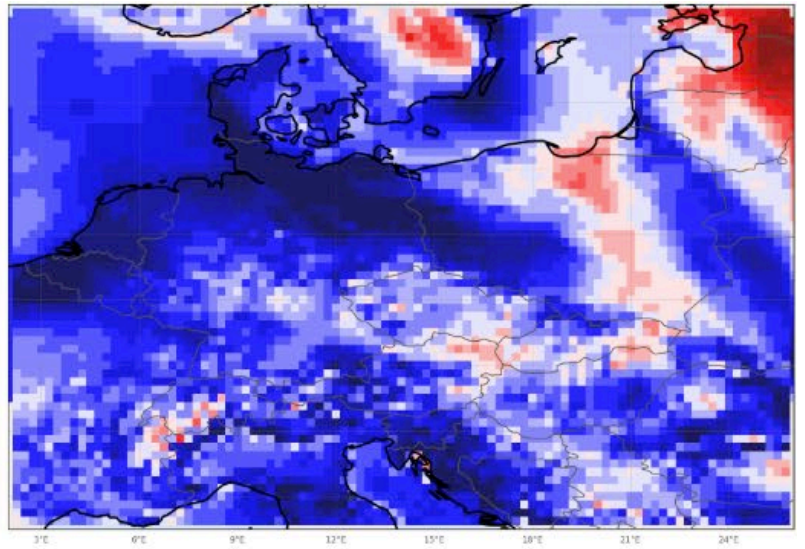
IFS

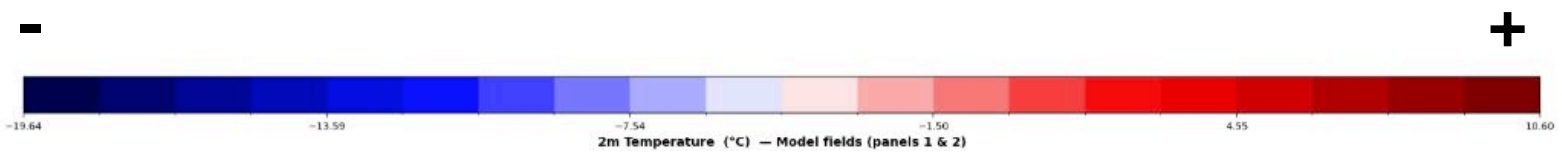


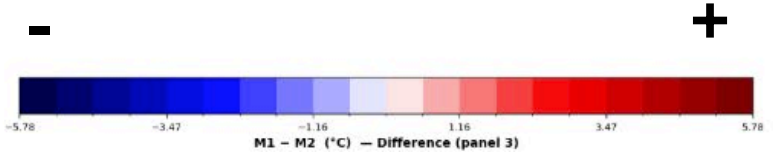
AIFS



IFS-AIFS

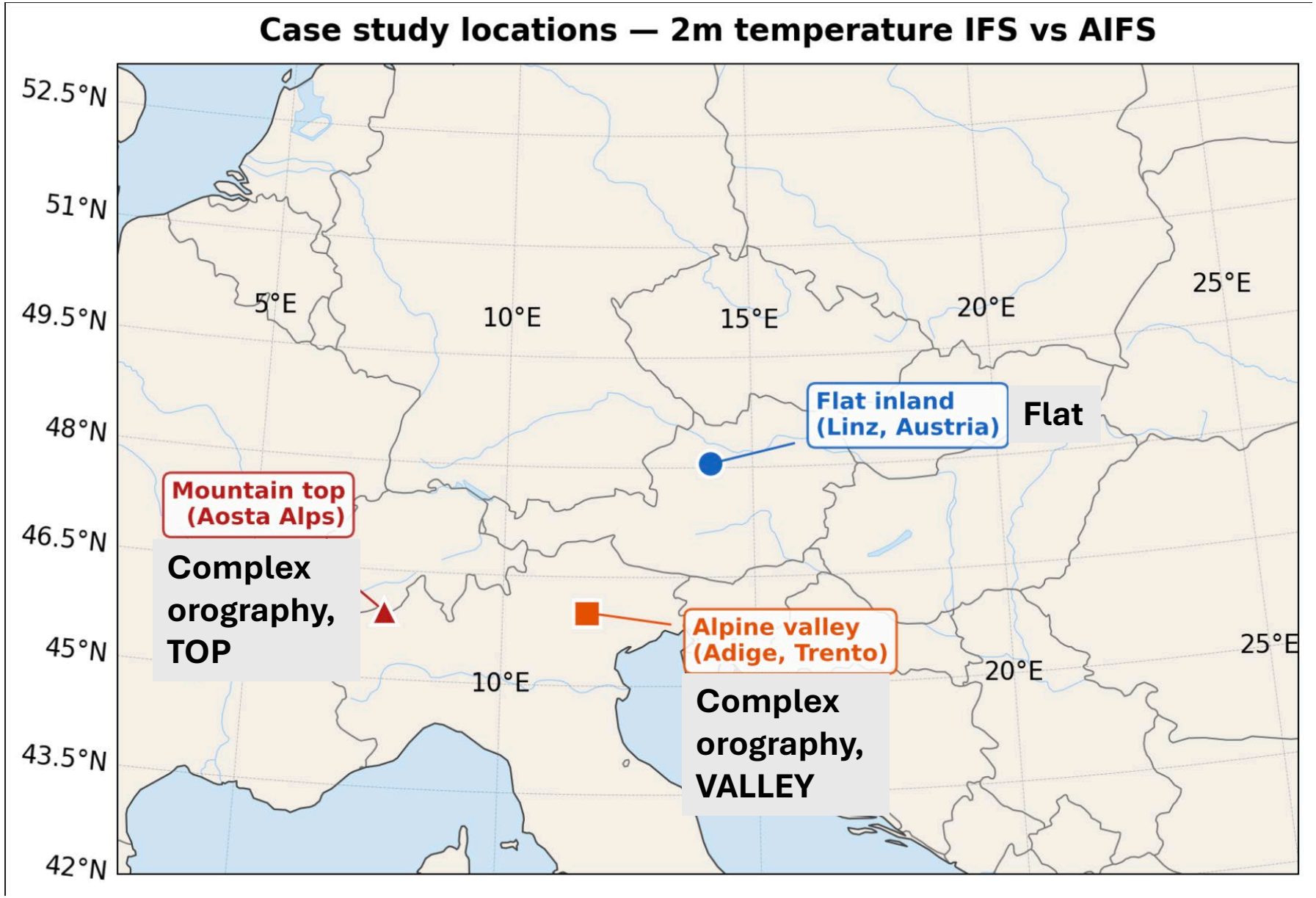


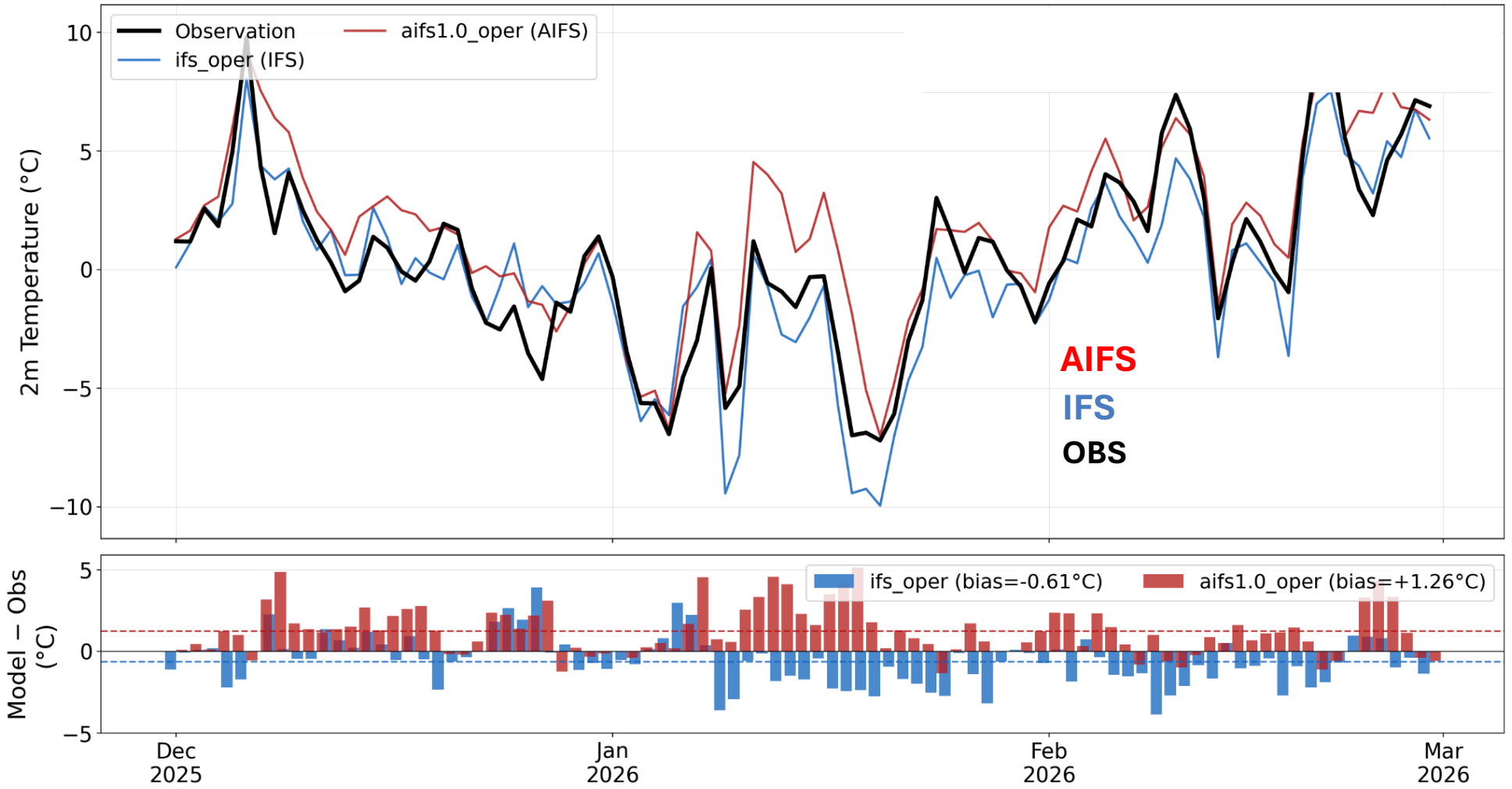




Both, **IFS** and **AIFS underestimate** the **cold extremes**

IFS generally **colder** than AIFS


Case study locations — 2m temperature IFS vs AIFS
52.5°N
51°N
49.5°N
48°N
46.5°N
45°N
43.5°N
42°N
5°E
10°E
15°E
20°E
25°E
Flat inland
(Linz, Austria)
Flat
Mountain top
(Aosta Alps)
Complex
orography,
TOP
Alpine valley
(Adige, Trento)
Complex
orography,
VALLEY



Flat

winter day 3

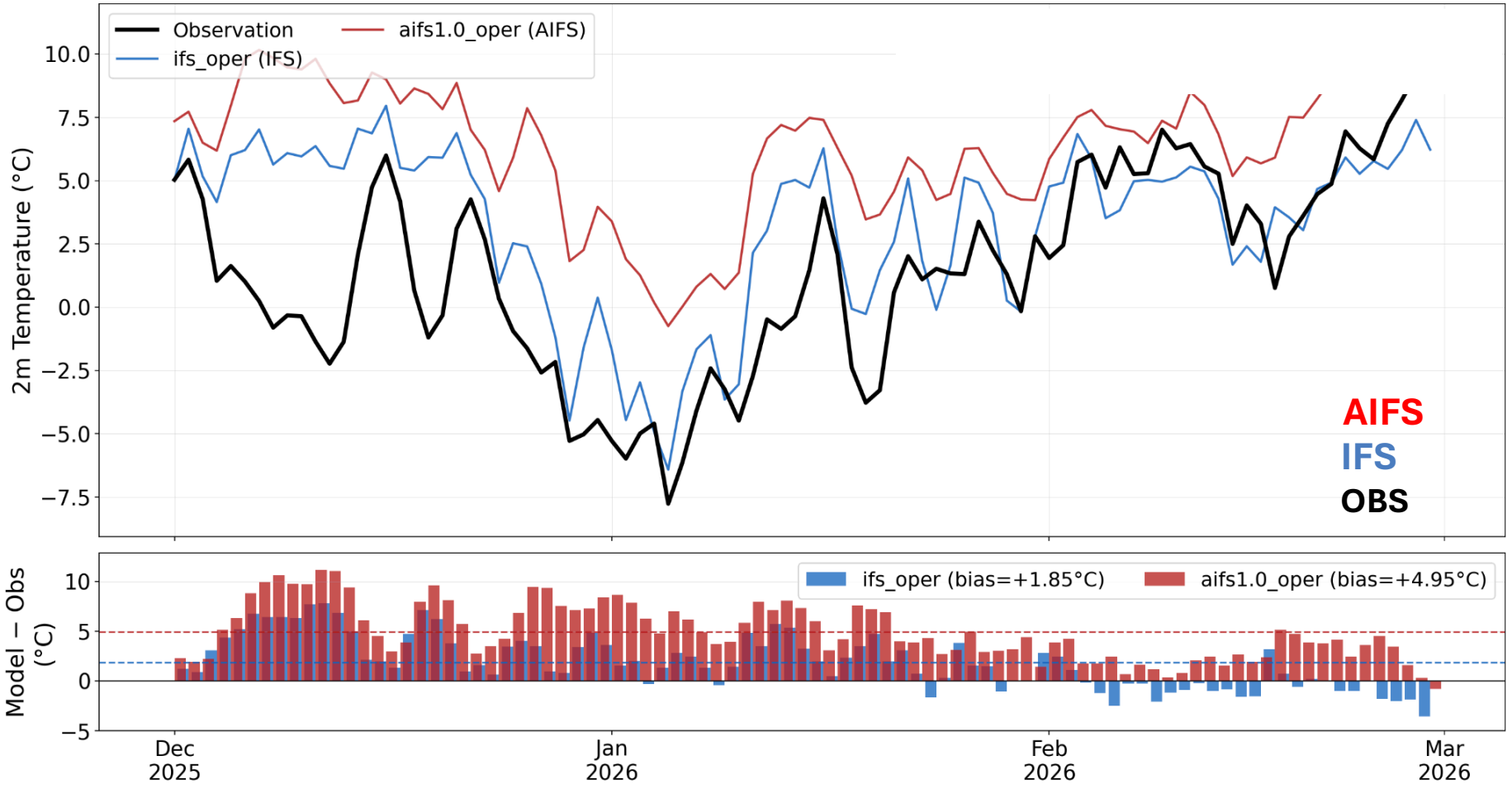
383 m

bias

AIFS warmer (+1.26 C)

IFS colder (-0.61 C)

Why does AIFS perform worse in this valley station?



Complex: VALLEY

winter day 3

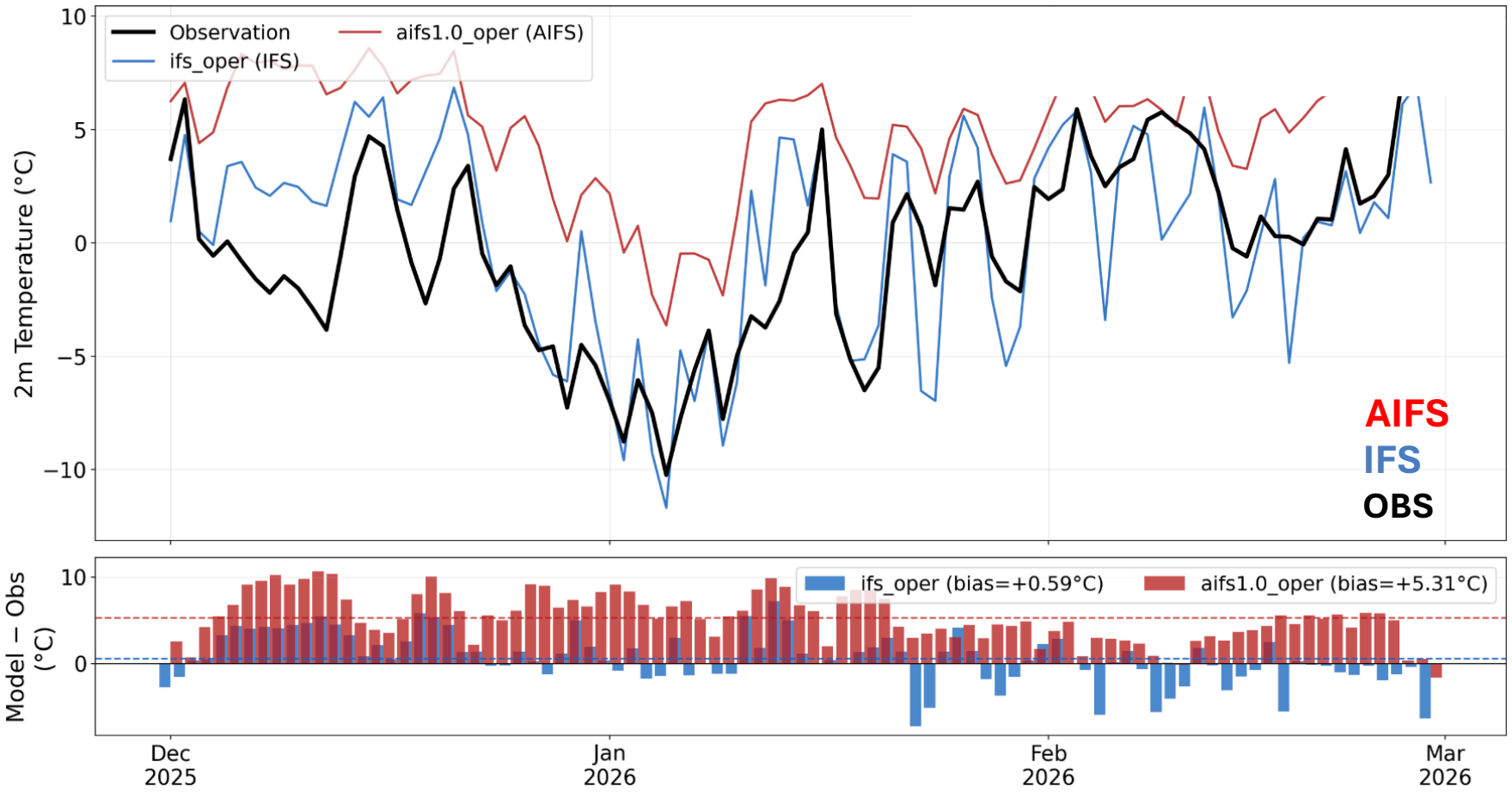
283 m

bias

AIFS much warmer (+4.95 C)

IFS warmer (+1.85 C)

The valley cold pooling is completely missed by AIFS



Complex: VALLEY

winter day 3

283 m

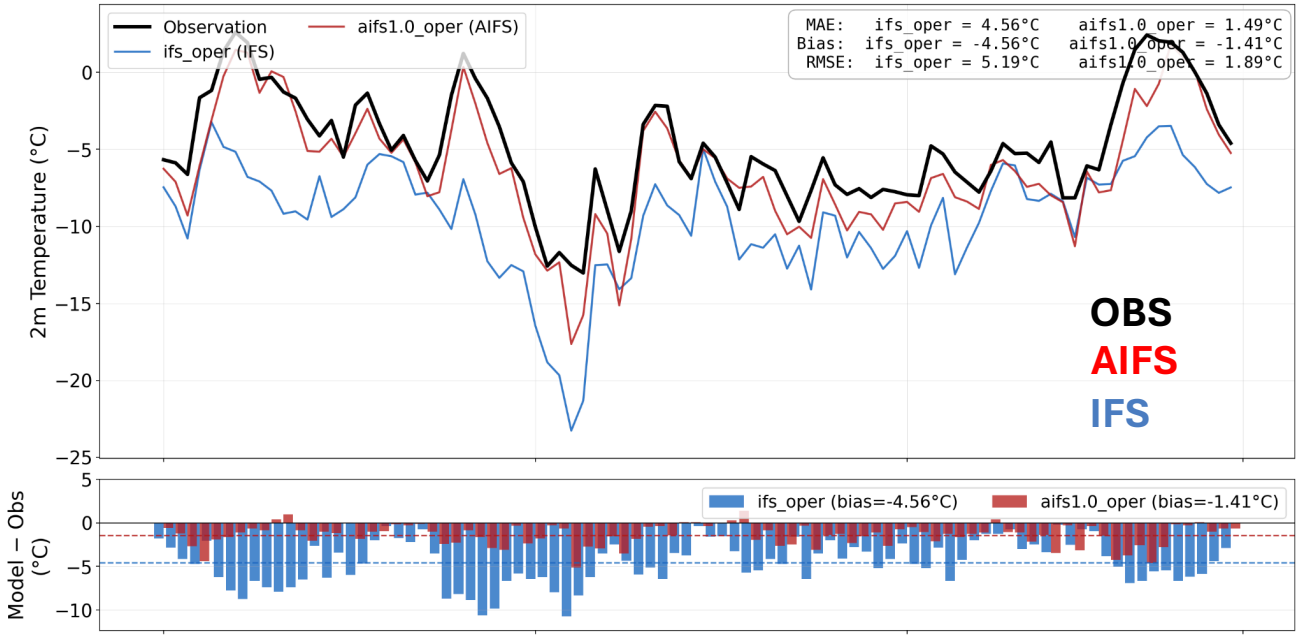
NIGHT 00 UTC

bias

AIFS much warmer (+5.31 C)

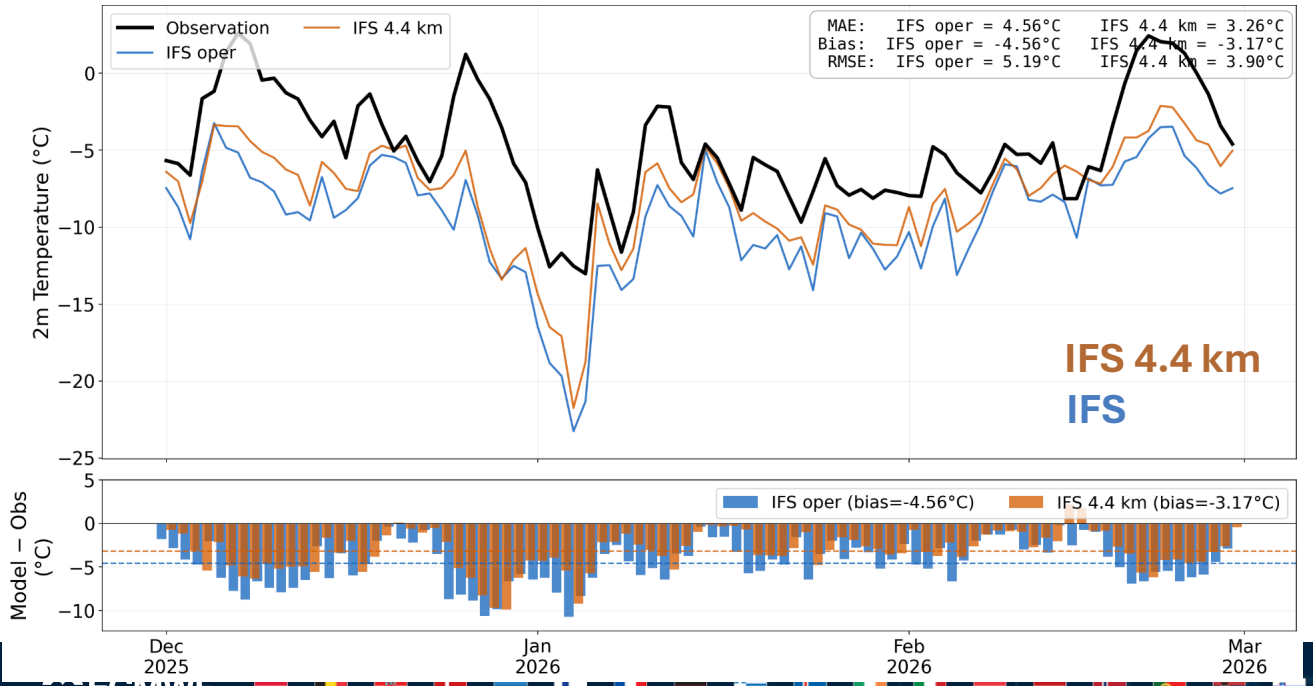
IFS warmer (+0.59 C)

NOTE: IFS and IFS 4.4 km look pretty similar for valleys, but 4.4 is slightly warmer (more mixing)



AIFS colder (-1.41 C)

IFS much colder (-4.56 C)



IFS 4.4 km colder (-3.17 C)

IFS much colder (-4.56 C)

Complex: TOP

winter day 3

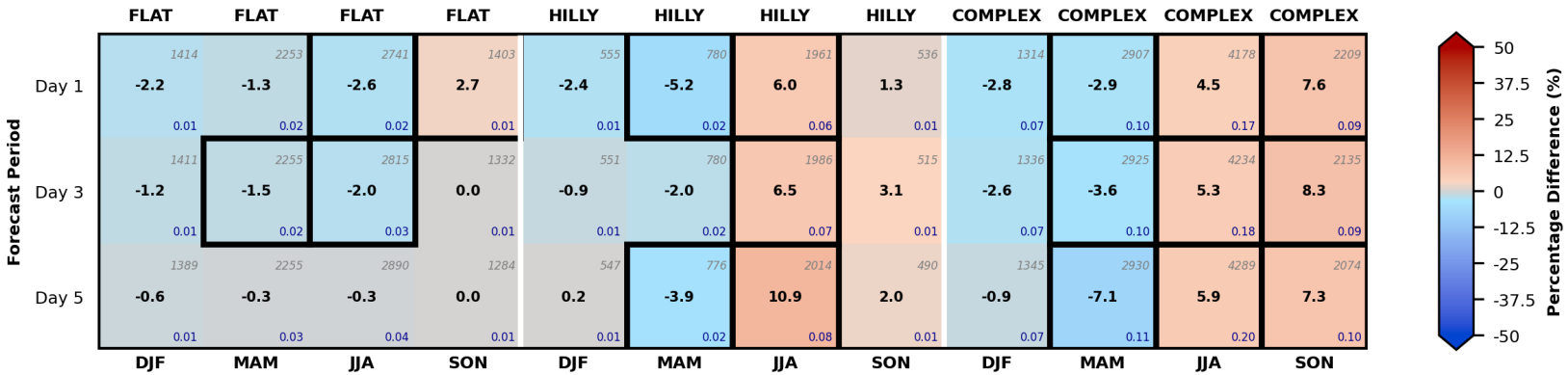
2511 m

For mountain tops, AIFS clearly outperforms IFS and IFS 4.4 km

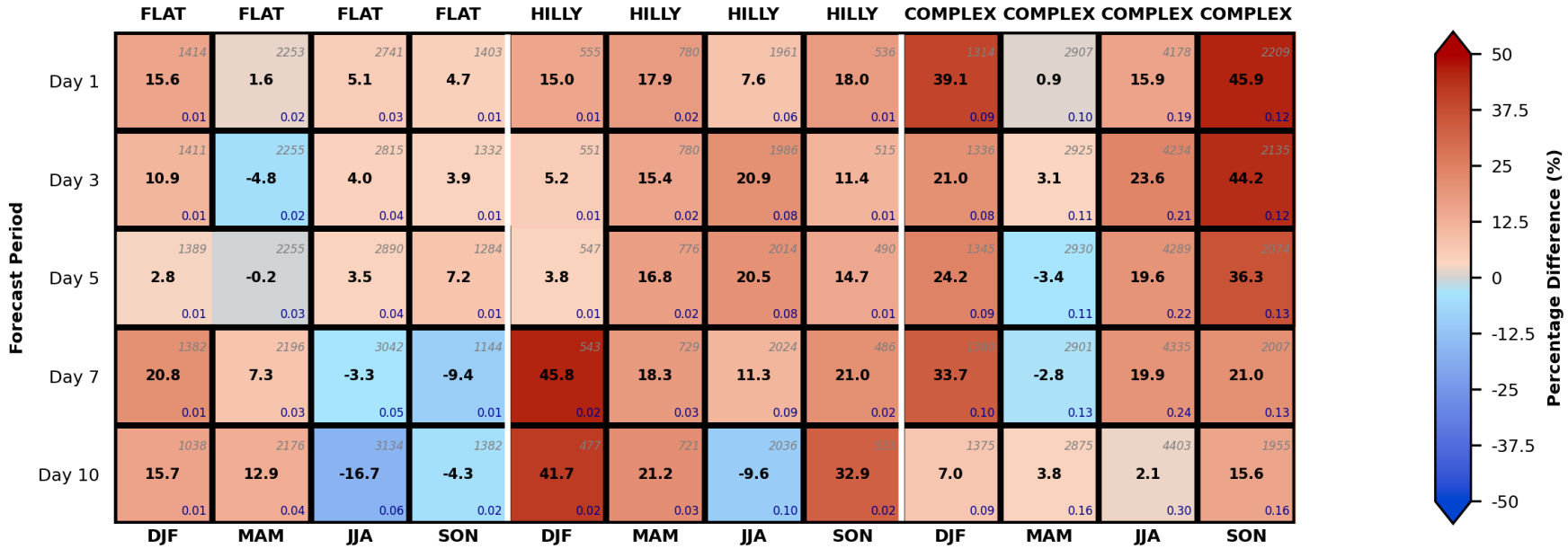
2 m temperature cold extremes (99th percentile station-based)

twMAE in Europe

IFS 4.4 km vs. **IFS**
(ref)



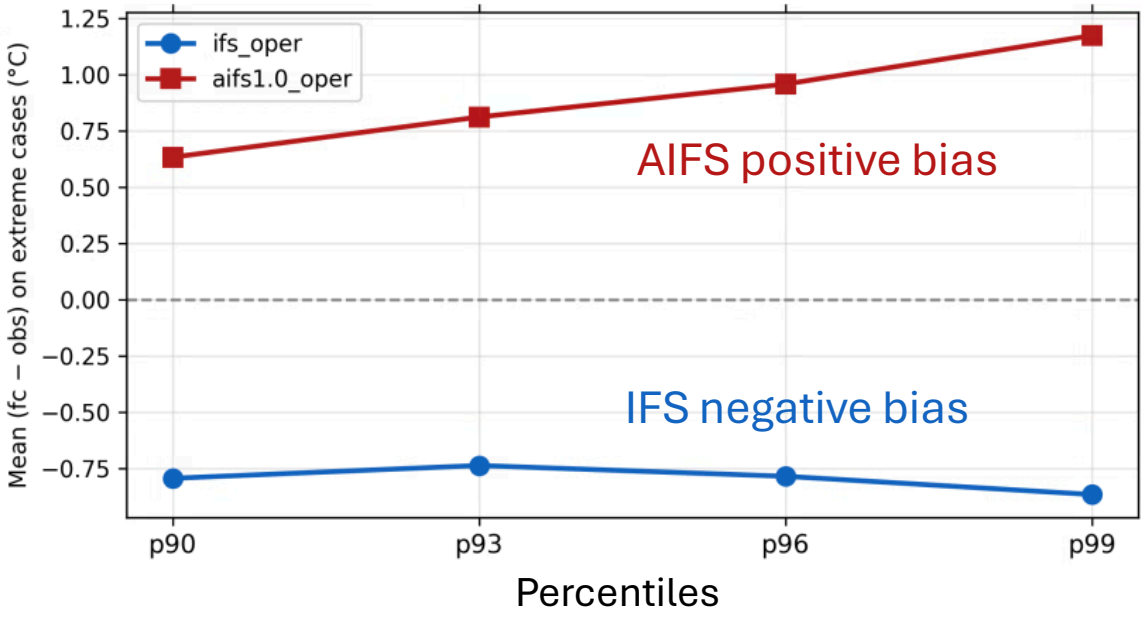
AIFS vs. **IFS**
(ref)

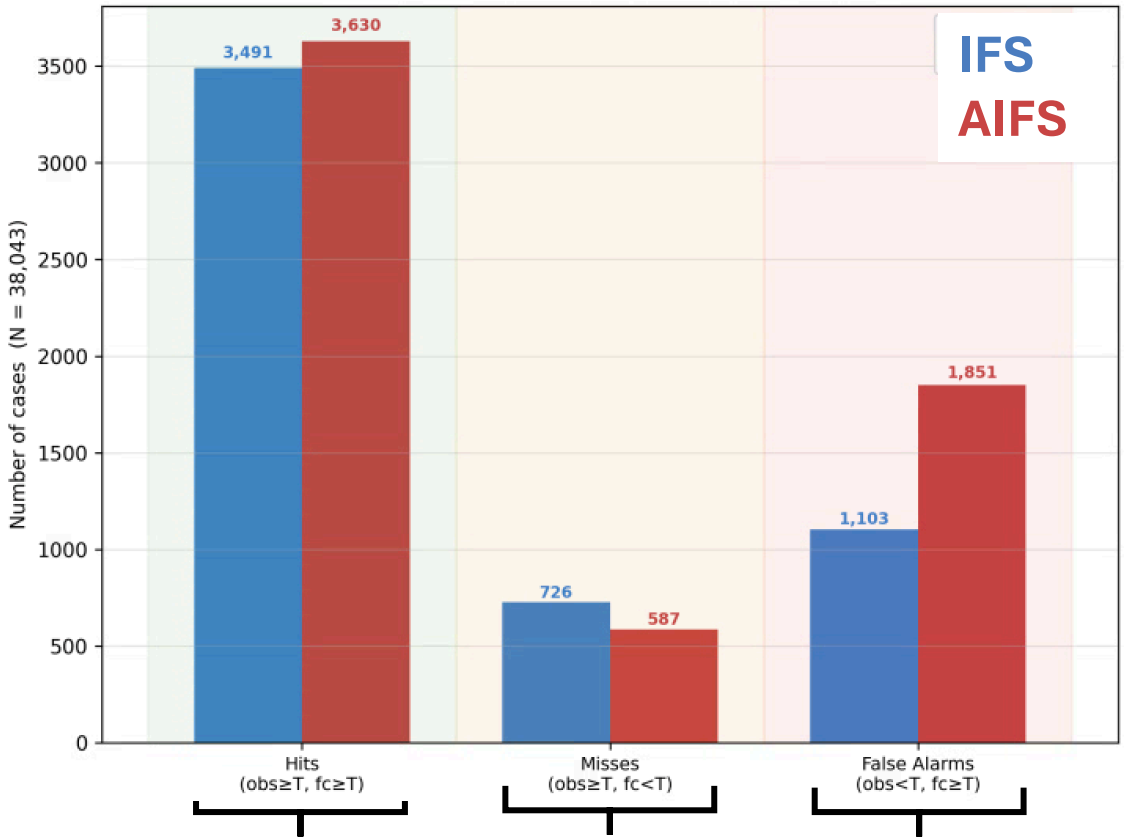


Complex orographic areas

summer day 3

Bias on warm extremes extremes





AIFS *slightly more hits*

AIFS *slightly fewer misses*

...but **AIFS** ***many more false alarms!!***

“Semi-subjective” veredict:

AIFS and IFS 4.4 km vs IFS oper — 2m temperature extremes			
		Flat terrain	Complex terrain
AIFS	❄	↑↑ much better (~20–35%)	↑↑↑ much better (~30–60%)*
AIFS	☀	↓ worse (~10–15%)†	↓↓ worse (~20–45%)†
IFS 4.4 km	❄	↑ better (~5–10%)	↑↑ much better (~20–50%)
IFS 4.4 km	☀	→ neutral / small gain	→ neutral / small gain

But very good on tops and bad in valleys at night

Warm bias!! Many false alarms

Resolving better physical processes (but too much mixing in valleys under stable conditions)

Thank you!

www.ecmwf.int