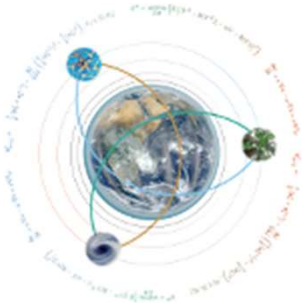


Méthodologie statistique pour améliorer la description de processus de petite échelle.

Exemple des ondes internes de gravité.

Groupe de Travail MathsInFluids, ENS Lyon

Aurélie FISCHER

LPSM, Université Paris Cité

Collaboration with :

Milena CORCOS (LMD)

Sothea HAS (LMD & LPSM)

François LOTT (LMD)

Riwal PLOUGONVEN (LMD)

Albert HERTZOG (LMD)

Aurélien PODGLAJEN (LMD)

Raj RANI (LMD)

IMPT project (Institut des Mathématiques pour la Planète Terre, CNRS)

DataWave : Collaborative Gravity Wave Research

VESRI project (Virtual Earth System Research Institute, Schmidt Futures)

Outline

- A few introductory facts on Statistical Modeling
- A short example : downscaling for wind speed
- Gravity waves and parameterizations
- Various uses in the field of Statistics / Machine Learning
- Our framework ERA5 & balloons
- Statistical learning framework
- Some results on the 2019 campaign
- Aggregation : learning on 2021 campaign & parameterizations

A few introductory facts on Statistical Modeling

Statistical Modeling

- Some natural phenomena may be too complex or too noisy to be described adequately by analytical equations alone.
- Nondeterministic models, based on observations : deterministic part + stochastic component.
- Small-scale processes : benefit from local observations.

Some possible features of statistical methods

- Parametric : simple to very large…

or

Nonparametric methods : flexibility, hyperparameters…

- Underfitting and Overfitting
- Interpretability vs black box : parametric, graphical aspect, variable importance…
- Computing time

- Several trade-offs to be found.

A short example : downscaling for wind speed

Example : Downscaling for wind speed

For a given location,

using past observations and past forecasts,

can we improve forecasts for a small-scale variable ?

Example : Prediction of wind energy production at a given location.

Data considered :
True observations : one specific location + stations over France
Model data : Numerical Weather Prediction Model outputs from the European Center for Medium-Range Weather Forecasts (ECMWF)

ECMWF

SIRTA Data

Alonzo et al., 2018

« Site Instrumental de Recherche
par Télédétection Atmosphérique »

Atmospheric Remote Sensing
Instrumental Site

Palaiseau (49N, 2E)
20 km south of Paris (France)
Semi-urban environment

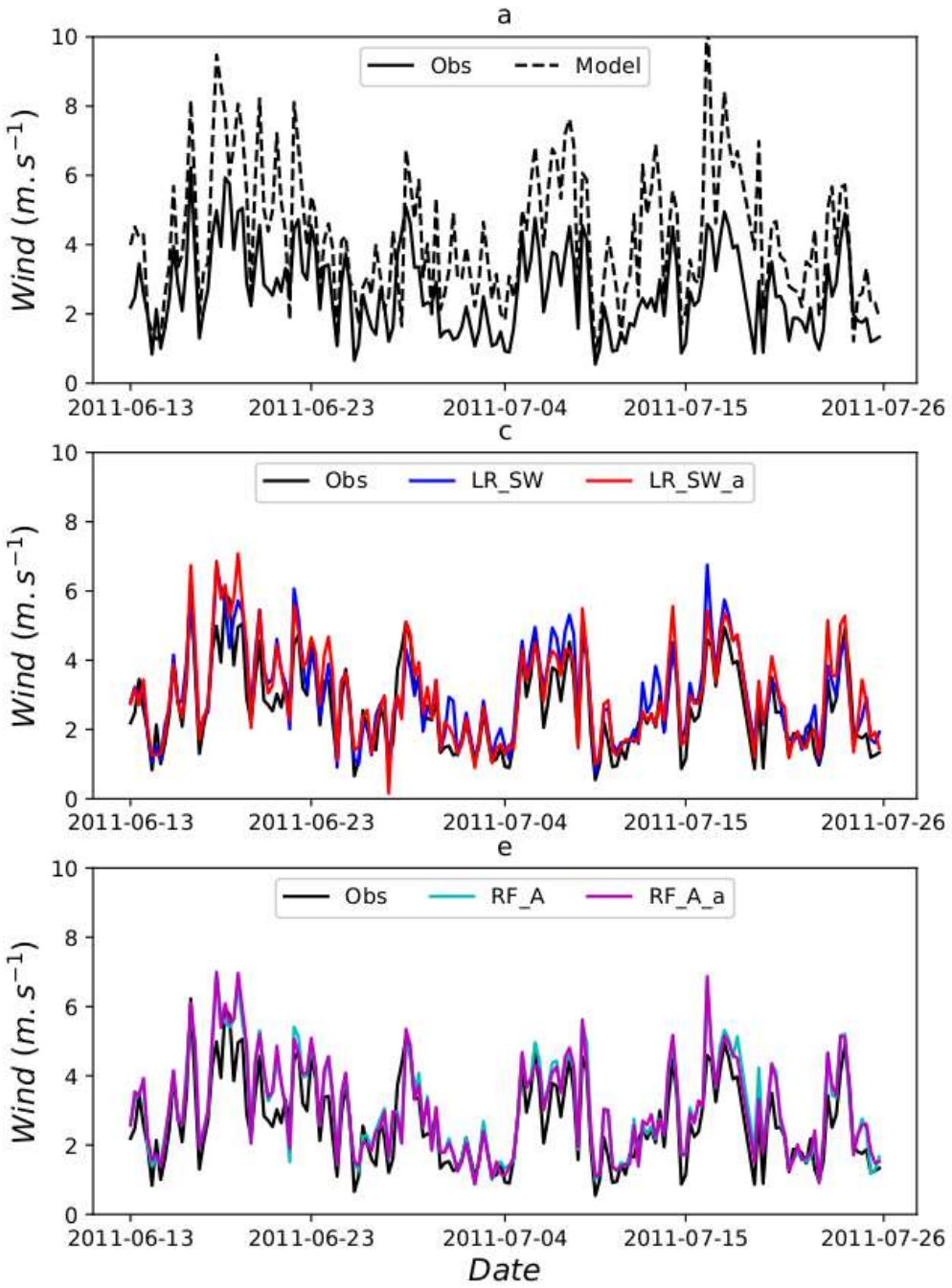
SIRTA Data

Variable of interest : wind speed at 10m and 100m.
→ Significant improvement for 10m wind speed.

Linear regression (parametric)

Random forests (nonparametric, based on trees)

Study of the relevance of the different explanatory variables



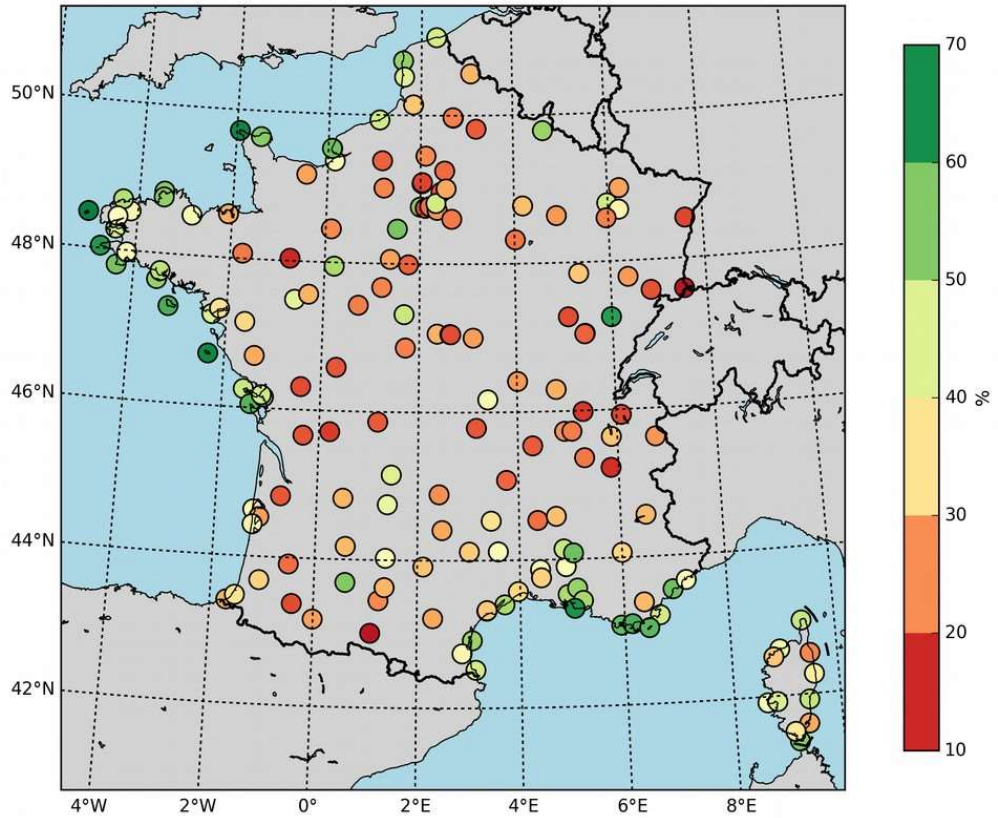
Data over France

Goutham et al., 2021

171 locations

Results :

- Comparisons of methods
- Study of relevant variables
- Geographical pattern



Reduction of RMSE for 10m wind, in %

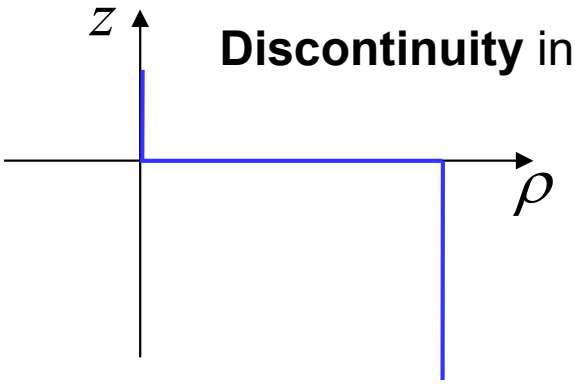
→ big improvement in coastal stations

→ moderate inland

Gravity waves and parameterizations

Gravity waves

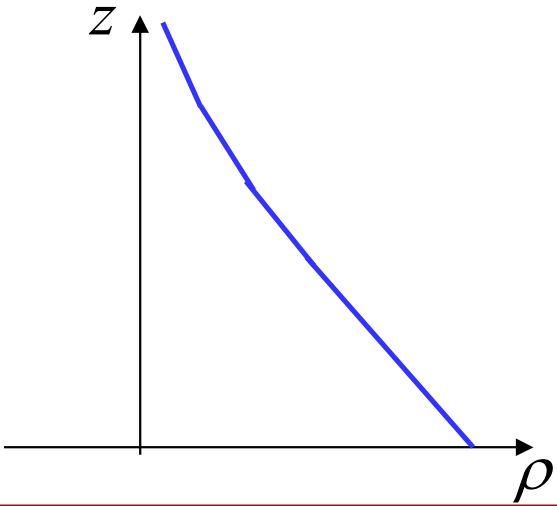
Waves due to gravity and to a contrast in density ρ in the vertical (denser fluid below...)



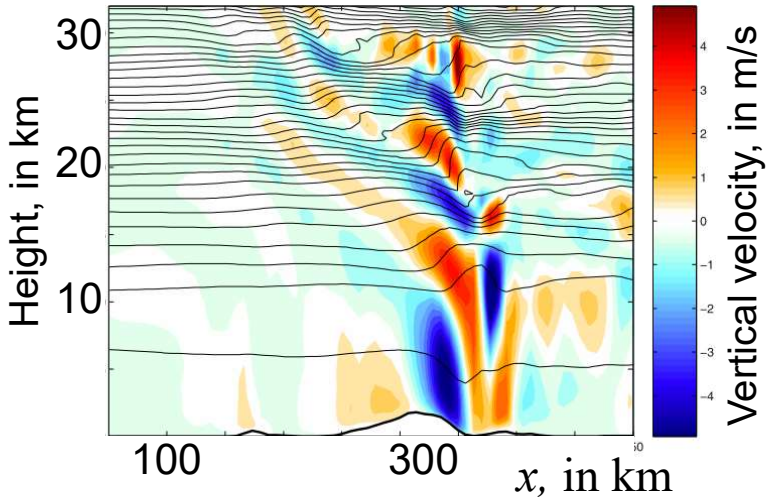
Discontinuity in density →

Periods starting from a few seconds

Continuous decrease in density →



Internal gravity waves



Periods from 5min to ~1-2 days

Gravity waves

Air displaced in the vertical :

- due to mountains (orographic waves)
- by jet streaks, fronts, convection

Impact for the general circulation : vertical transfer of momentum from the troposphere to the stratosphere and mesosphere.

Important role in daily weather + long-term climate fluctuations.

→ One of the wave families forcing the Quasi-Biennal Oscillation (QBO)

Quantity of interest = GW momentum fluxes
accelerate or decelerate flow higher up = change in air momentum

Need of parameterizations

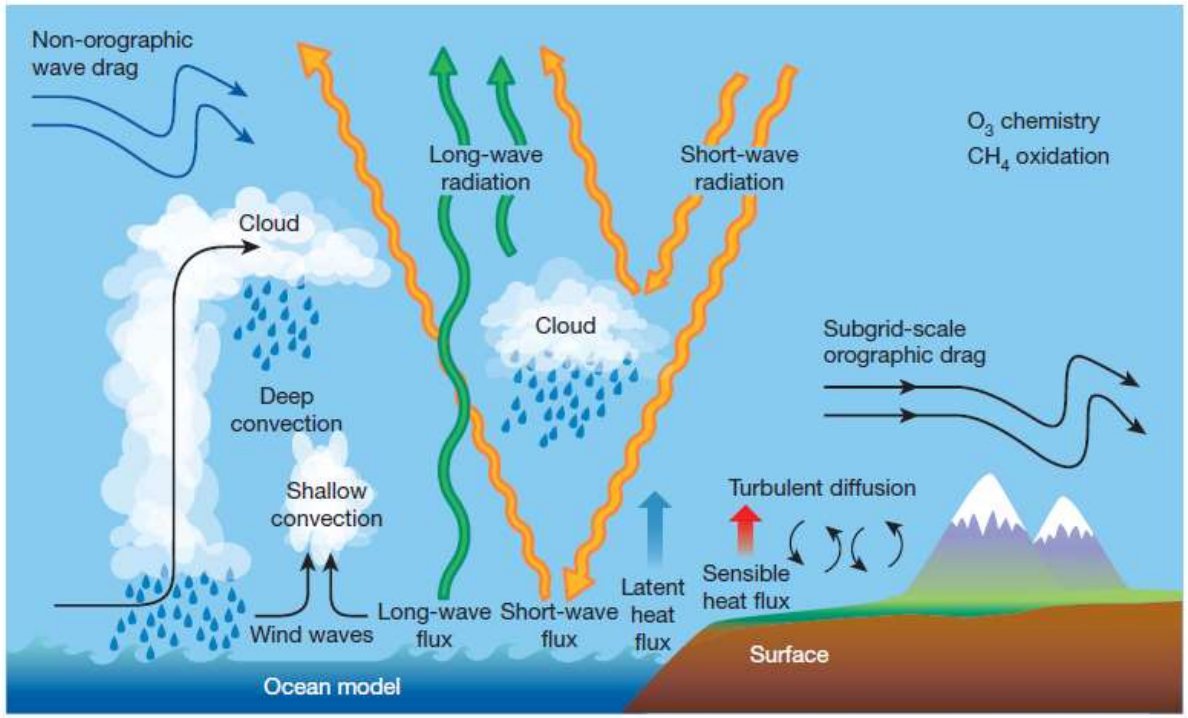


Figure 2 | Physical processes of importance to weather prediction. These are not explicitly resolved in current NWP models but they are represented via parameterizations describing their contributions to the resolved scales in terms of mass, momentum and heat transfers.

Bauer et al 2015

Need of parameterizations

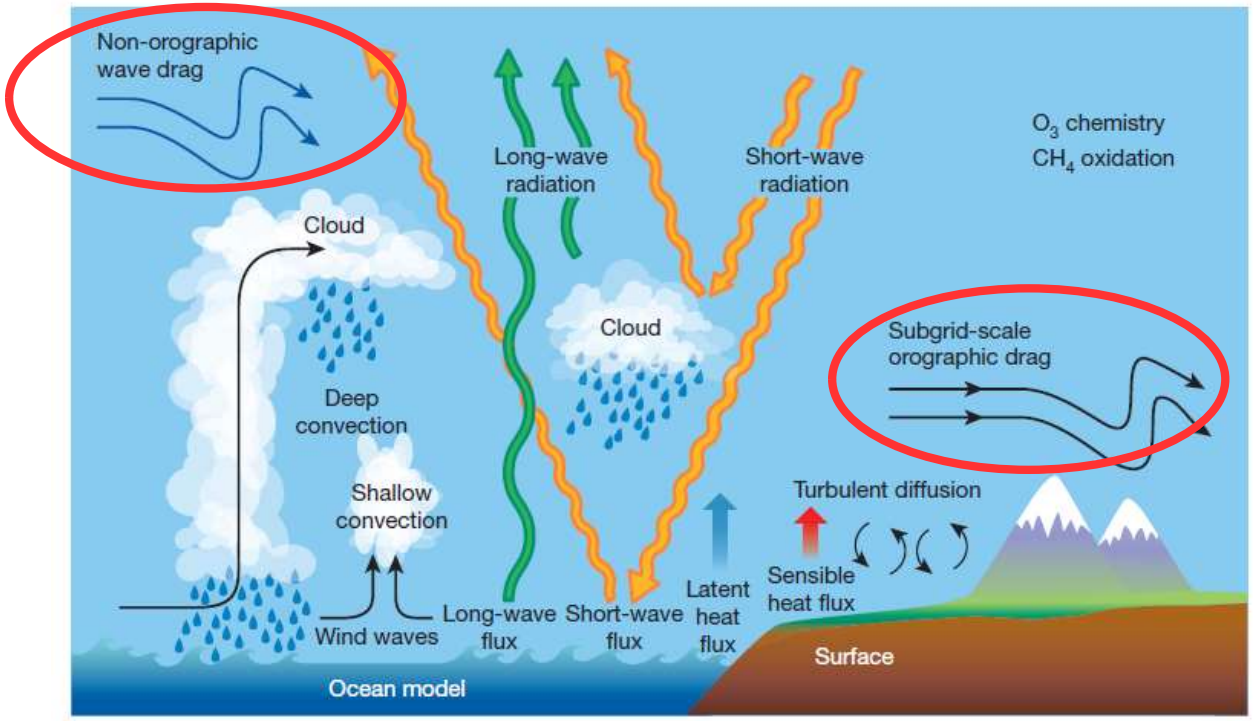


Figure 2 | Physical processes of importance to weather prediction. These are not explicitly resolved in current NWP models but they are represented via parameterizations describing their contributions to the resolved scales in terms of mass, momentum and heat transfers.

Bauer et al 2015

GW are subgrid scale, unresolved processes

→ necessary to parameterize GW

Parameterizations

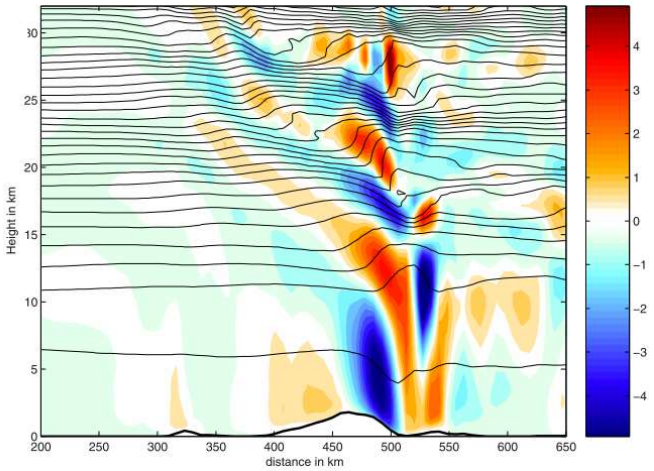
In climate and weather models, workaround to represent subgrid-scale = unresolved processes.

→ Even if unable to include GW in the model, using the knowledge of their actions, represent their impacts on the resolved flow.

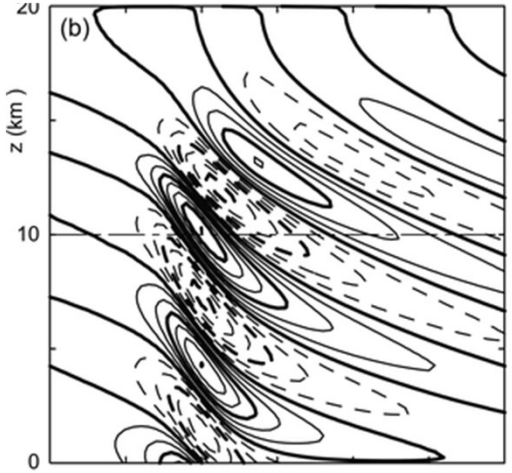
- Universal : in any location, relies on resolved physical variables, not location-specific.
- Physics-based : ideally, should be based on physical laws, as the equations of motions are for the resolved flow.

Example : orographic waves

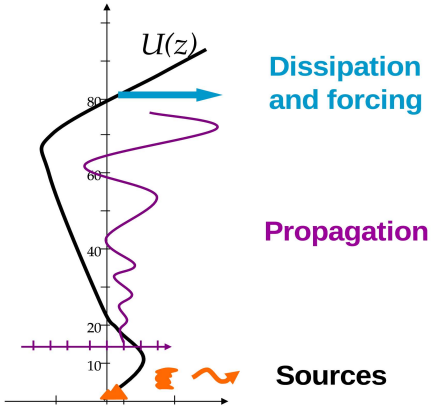
Real / realistic



Idealized, analytic



Parameterization



Parameterizations

GW dynamics simplified to minimum :
- source specification
- vertical propagation
- dissipation and forcing of the flow

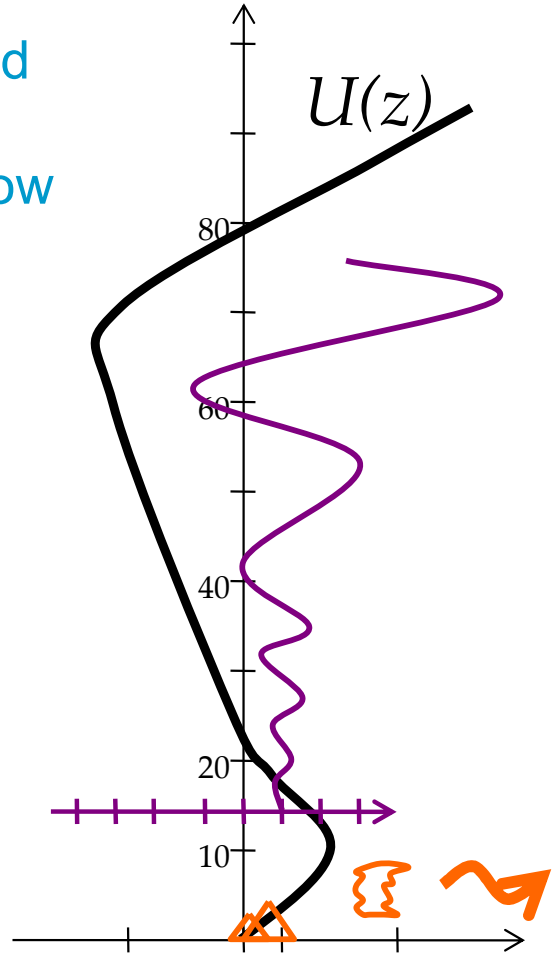
Much research targeting sources.
- Fairly arbitrary, poorly constrained.
- Parameters conveniently tuned.

→ Errors, uncertainty

Dissipation and forcing of the background flow

Propagation

Sources



Various uses in the field of Statistics / Machine Learning

Some examples of machine learning applications

Emulate parameterizations to save computing time

« Metamodel » from higher resolution simulations

Data-driven parameterizations built using machine learning

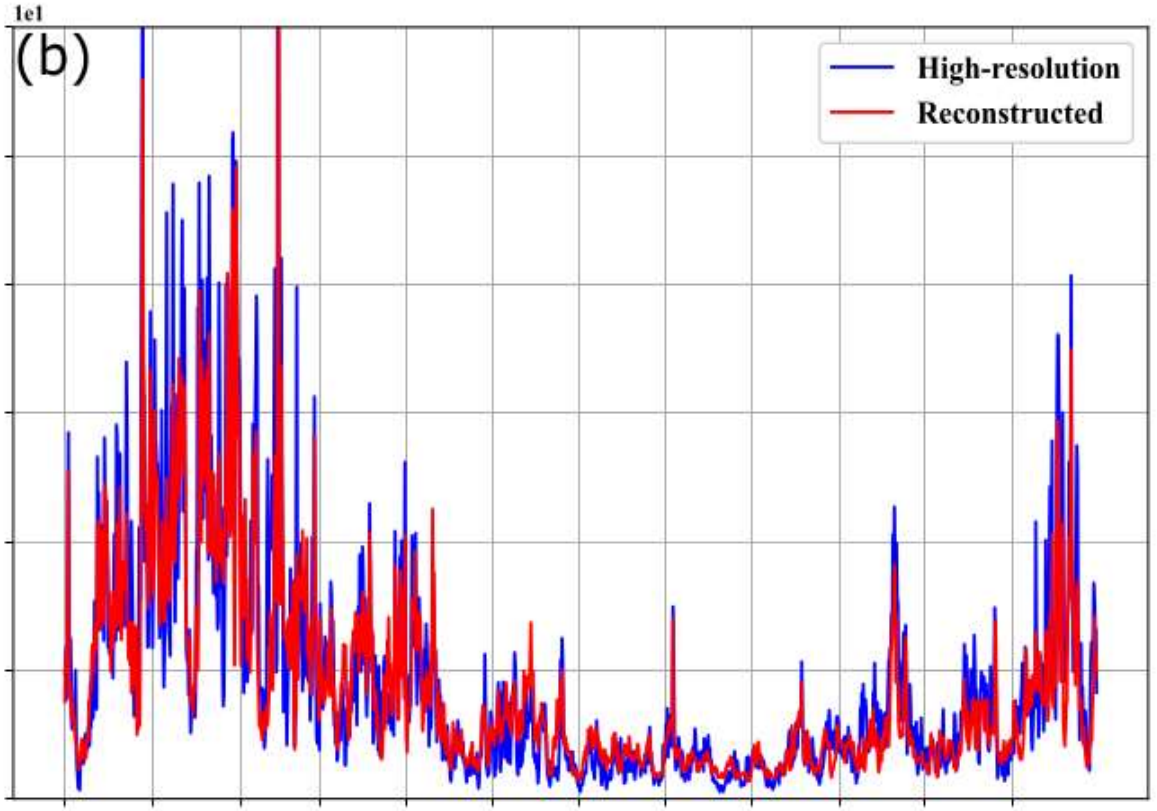
Relate large scale flow to local observations (gravity waves momentum fluxes)

Metamodeling

Amiramjadi et al, 2020

ECMWF moderate resolution → GW information (target = model information)
ECMWF low resolution → Large-scale flow

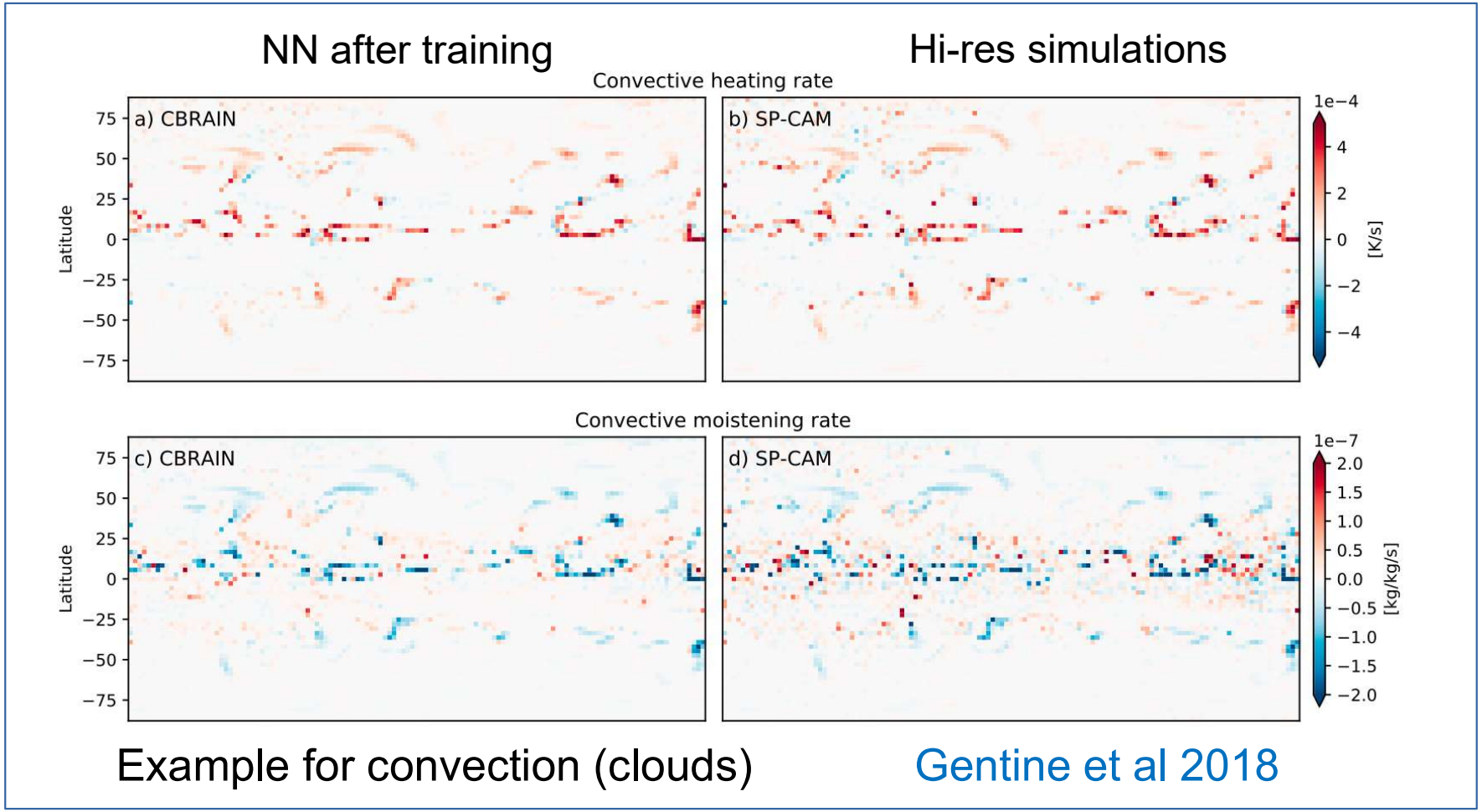
Random Forests to reconstruct GWs



Data-driven parameterizations using ML

For processes that are *resolvable* (gravity waves, clouds), short high-resolution simulations provide information.

→ Capture relationship between resolved and unresolved processes



Our framework ERA5 & balloons

ECMWF data → Information on the large-scale flow

Stratospheric balloons → Accurate observations on gravity waves

ML to reconstruct observed GW momentum fluxes from large-scale.

Observations : stratospheric balloons

Superpressure balloons, 11 and 13 m in diameter

Flight levels : ~18 and ~20 km Lifetime : 2 to 3 months

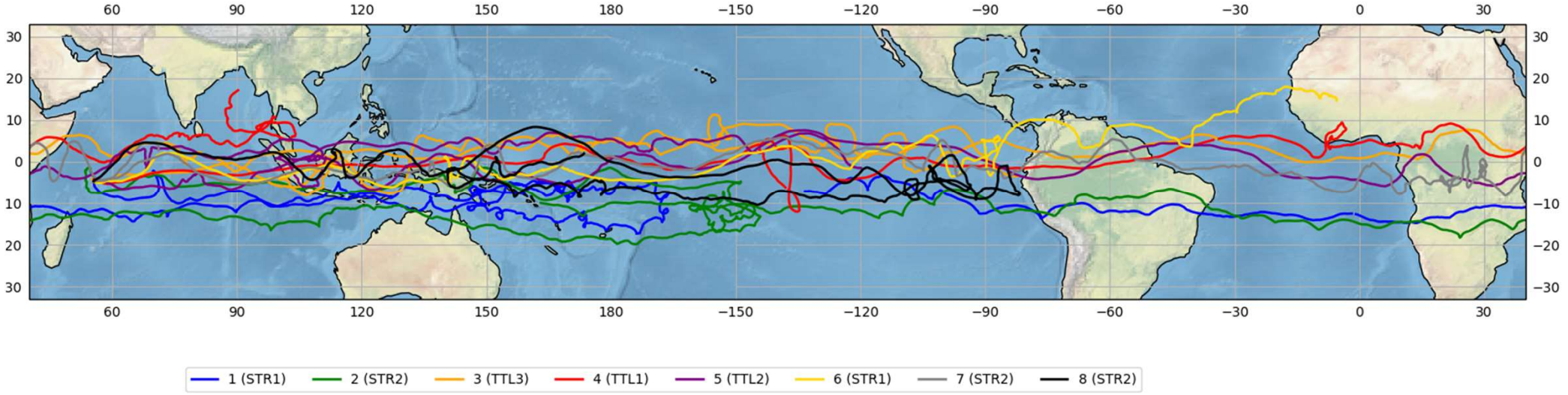
Observations : stratospheric balloons

Stratéole 2

French-US project

2019 campaign C0 :
8 balloons,
November 2019 to February 2020, along the tropics,
680 days of measurements.

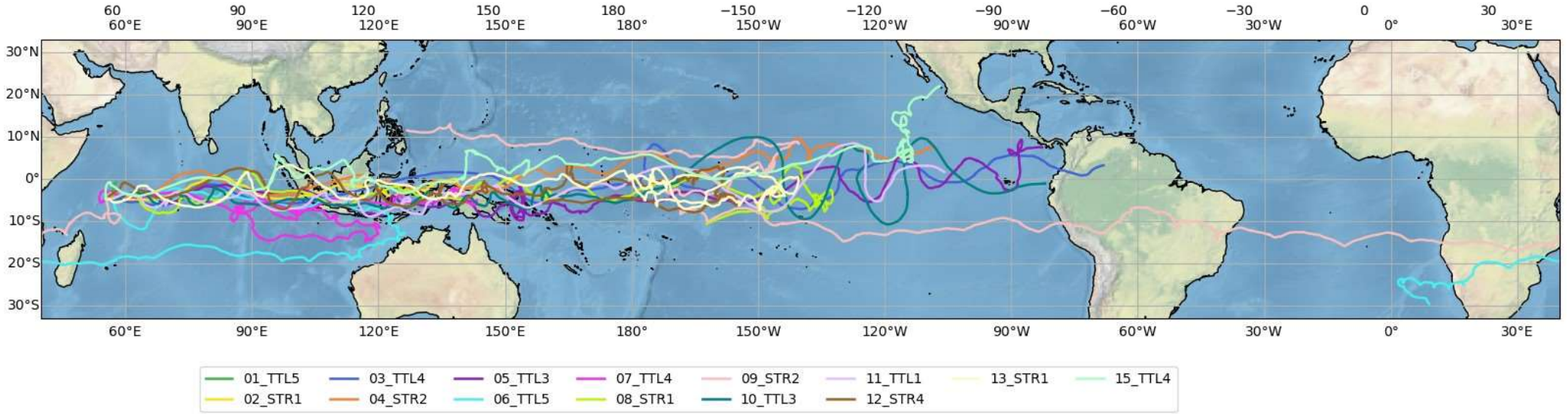
Data registered = 30s observations of position (wind), in-situ air pressure + temperature.



Observations : stratospheric balloons

Stratéole 2

2021 campaign C1 :
17 balloons
October 2021 to January 2022.



Observations : stratospheric balloons

Unique and valuable source of information on GW :

Quasi-Lagrangian behavior.
→ Direct access to the intrinsic frequency of the GW, thanks to in situ measurements
(not remote sensing, from temperature data : incertainty).

→ accurate estimate of key quantities, using wavelet analysis : momentum fluxes (Hertzog et al., 2012).

→ Large spatial cover since the balloons drift.

Some remarks on the balloons

Direction :

Surface wind near the Equator has direction East→West

At balloon altitude, winds alternate between westerlies and easterlies, period of ~28 months (QBO) → East

+ reversal

+ 2 balloons → West (further from the Equator, in South hemisphere)

Oscillation period 3min : high frequency GW → period 15min

Explanative variables from Reanalysis ERA5

5th generation of the European Reanalysis

Reanalysis : historical observations + numerical models
→ weather/climate datasets
ERA5 : state-of-the-art global atmospheric reanalysis dataset (hourly from 1 to 137 vertical levels).

Extracted variables : precipitation, pressure, wind and temperature profile (67 vertical levels) at 5 x 5 horizontal grid points of 1° x 1° (100km) resolution.

Question : Which large-scale variables are most informative about GW ?

ECMWF

Explanative variables from Reanalysis ERA5

Inputs :

Temperature : temp

Zonal and meridional wind : u and v

4 levels : 19, 9, 2km and surface level (0km).

log surface pressure: lnsp

Solar zenith angle : sza

Precipitation: tp, tp_{mean}, tp_{sd}

Targets : two types of absolute, eastward and westward GWMFs

- High frequency waves (HF) : period 15mn to 1h.
- Wide frequency waves (WF) : period 15mn to 1 day.

Statistical learning framework

Statistical learning setting

We observe a sample $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ from a generic random pair (X, Y) taking its values in $\mathbb{R}^d \times \mathbb{R}$.

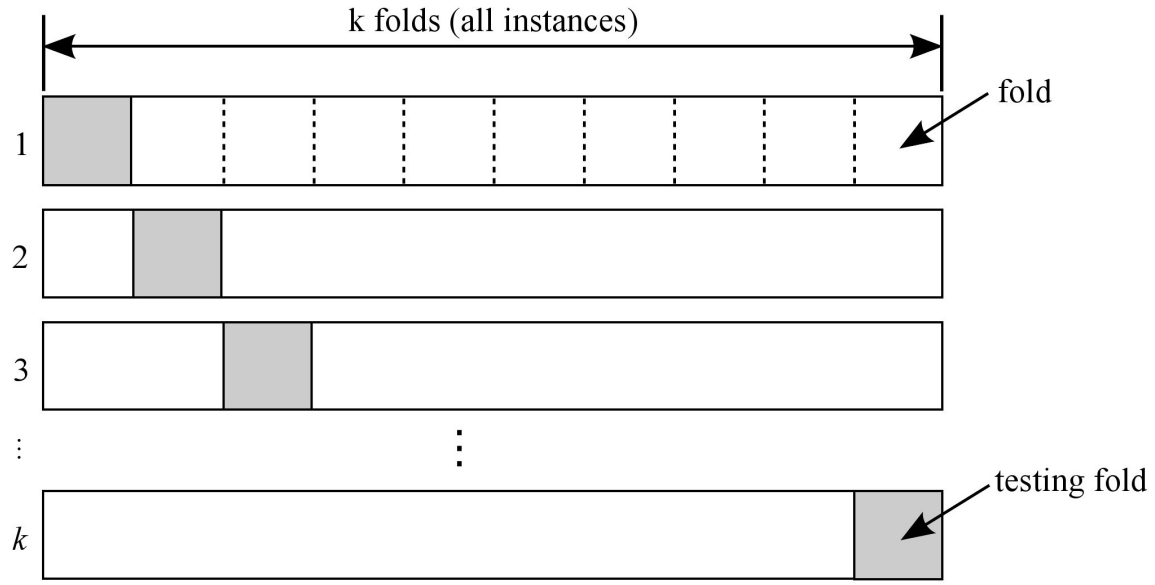
Explain the variable of interest / output Y using the different features or inputs $X = (X^1, \dots, X^d)$.

In other words, based on the data $\mathcal{D}_n$, we look for some function g such that $Y = g(X)$.

For new x, predict associated y by $g(x)$.

Cross-validation

Given a number K, we divide the sample into K blocks (for example, $K = 5$ or $K = 10$).



The K-fold cross-validation consists then in giving successively the status of validation sample to each block, the other blocks forming the learning sample.

Cross-validation algorithm

Random partition of $\{1, \dots, n\}$ into K subsets $I_1, \dots, I_K$ of similar sizes.

For $k = 1, \dots, K$,

$\mathcal{D}_{-k} = \{(X_i, Y_i), i \notin I_k\}$ learning set, $\mathcal{D}_k = \{(X_i, Y_i), i \in I_k\}$ validation set.

For a set of values of a parameter to be calibrated $\lambda \in \{\lambda_1, \dots, \lambda_d\}$,

construction of the learning method f_λ^{-k} on the set $\mathcal{D}_{-k}$ and computation

of the error on the set $\mathcal{D}_k$: $e_k(\lambda) = \sum_{i \in I_k} \mathcal{L}\left(Y_i, f_\lambda^{-k}(X_i)\right)$.

For every value of λ, average error on the K blocks.

Choose the parameter value λ minimizing the error.

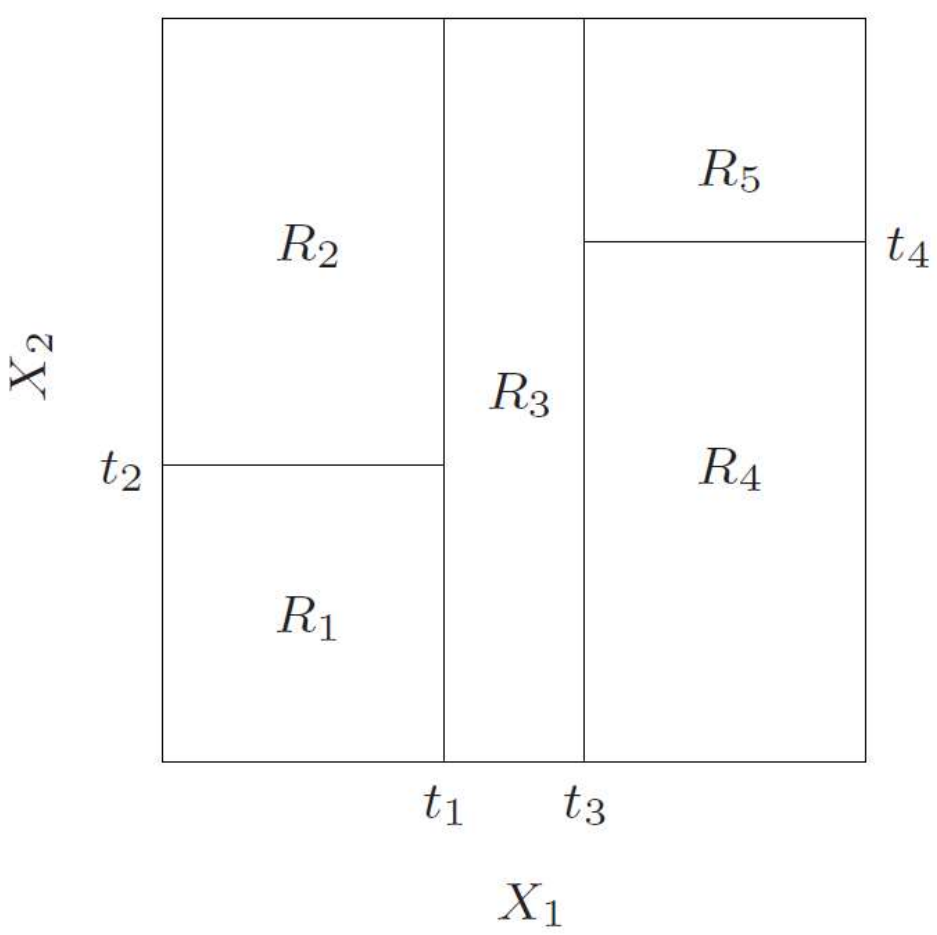
Statistical methods

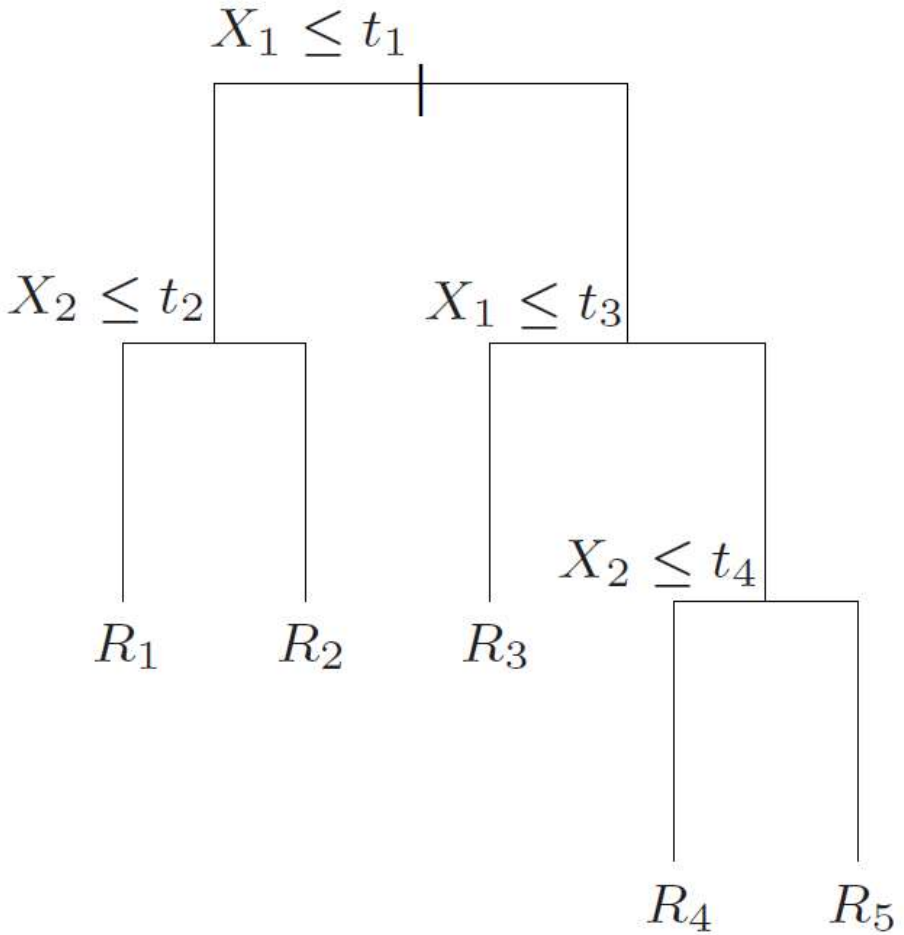
- Parametric :
  - Linear regression : $Y = X\beta + \varepsilon$.
  - Neural networks

- Nonparametric :
  - Neighborhood, proximity notion : nearest neighbor, kernel rule
  - Trees & aggregation

Statistical methods

Principle of a regression tree.





Statistical methods

Nonparametric tree-based methods : combining several regression trees.

- Random forests : bootstrap samples (resampling) = bagging + subset of variables, at random.
- ExtraTrees : initial sample, subset of split thresholds, at random.
- Boosting : weak estimators, iterative, based on weights.

Out of bag samples

For each observation (X_i, Y_i), on may construct the aggregated rule corresponding to the trees built on bootstrap samples in which this observation i does not appear.

OOB error $\sim$ Cross-validation

Variable importance

At each split in each tree, the improvement in the split-criterion is attributed to the splitting variable.

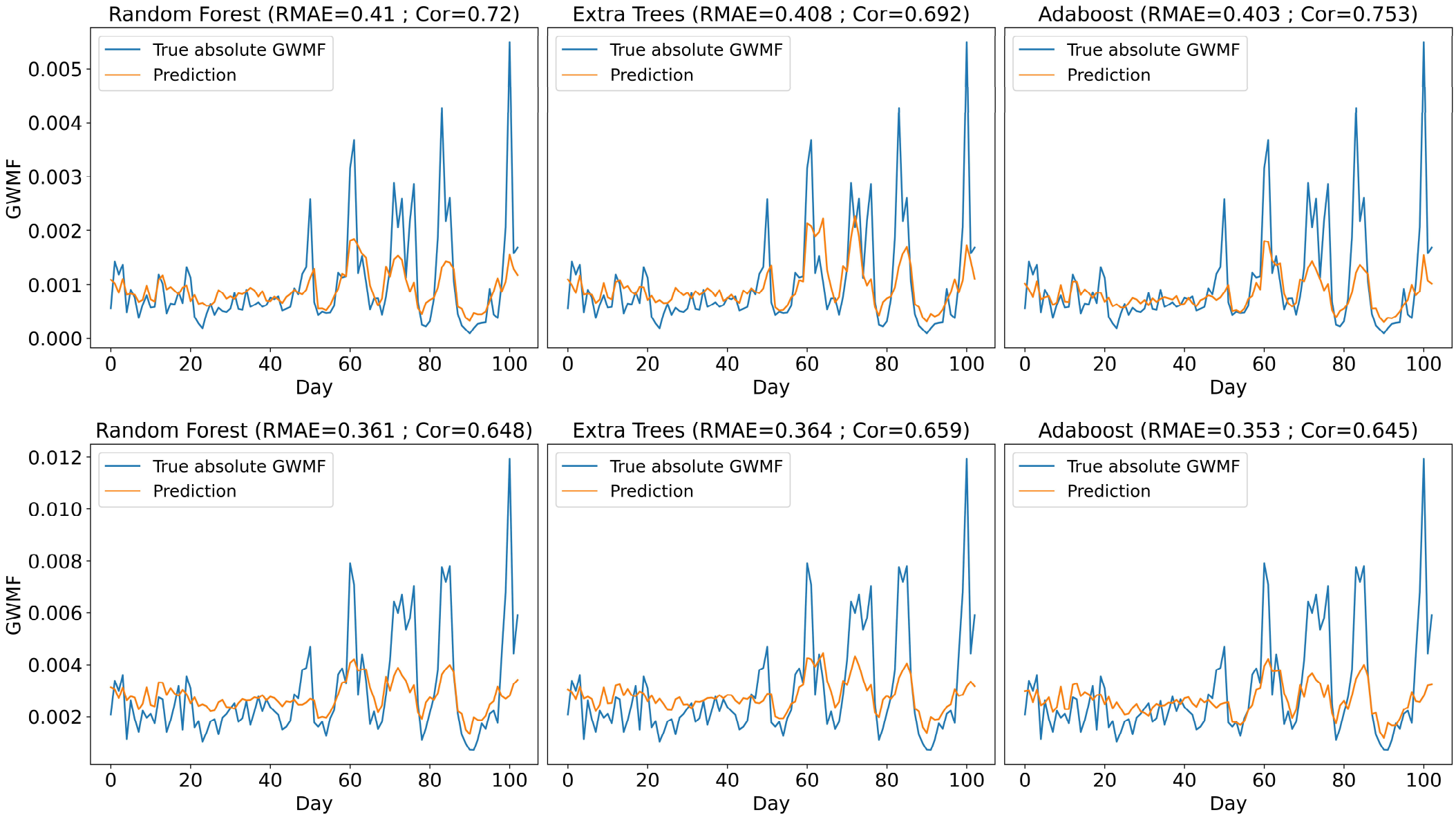
For each variable, values are accumulated over all trees in the forest.

When a tree is grown, the prediction accuracy is compared with the prediction accuracy when the values for the j-th variable are randomly permuted in the OOB samples.

The decrease in accuracy due to this permuting is averaged over all trees and used as a measure of the importance of variable j.

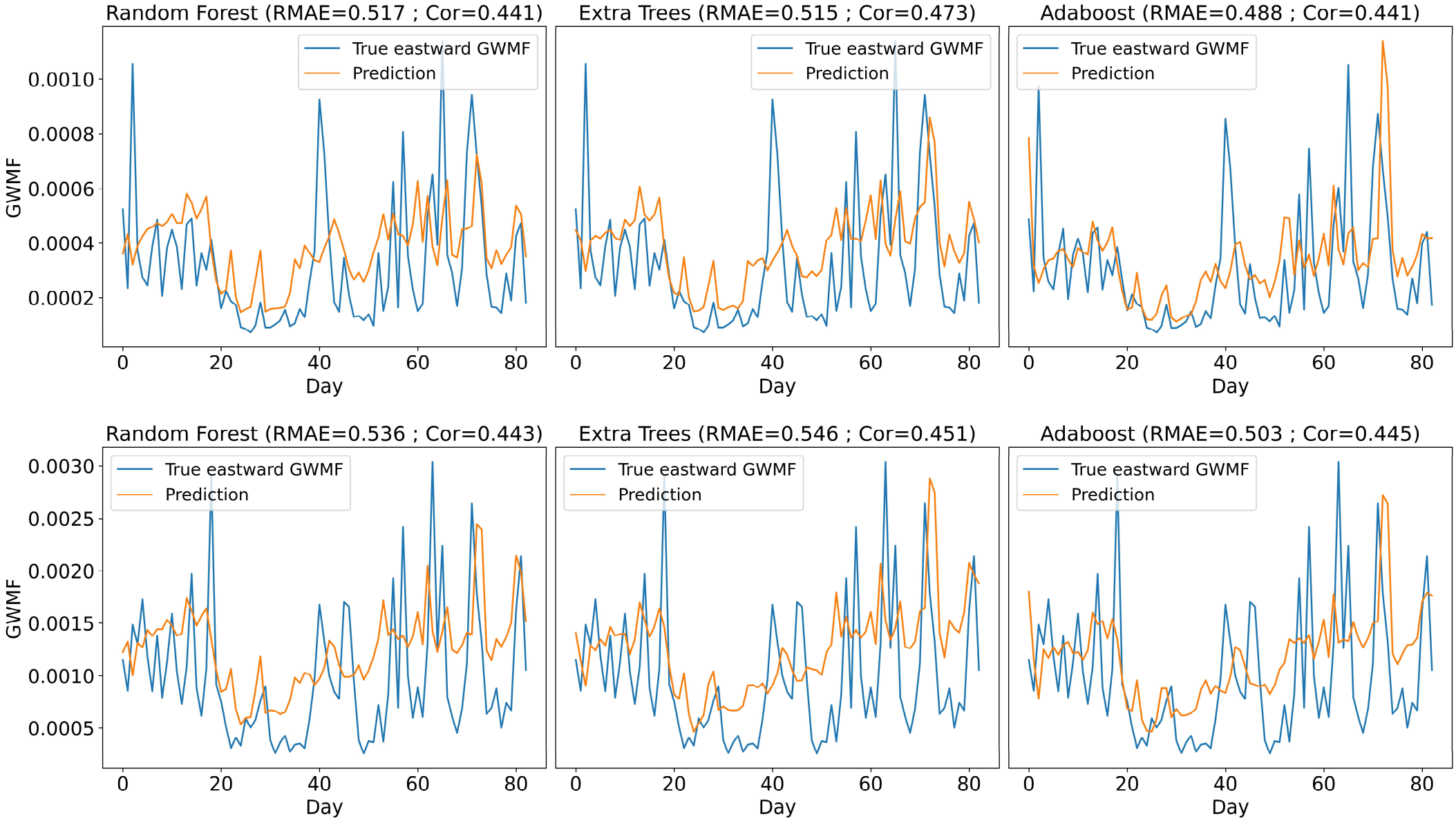
Some results on the 2019 campaign

Absolute GWMF : Balloon 2



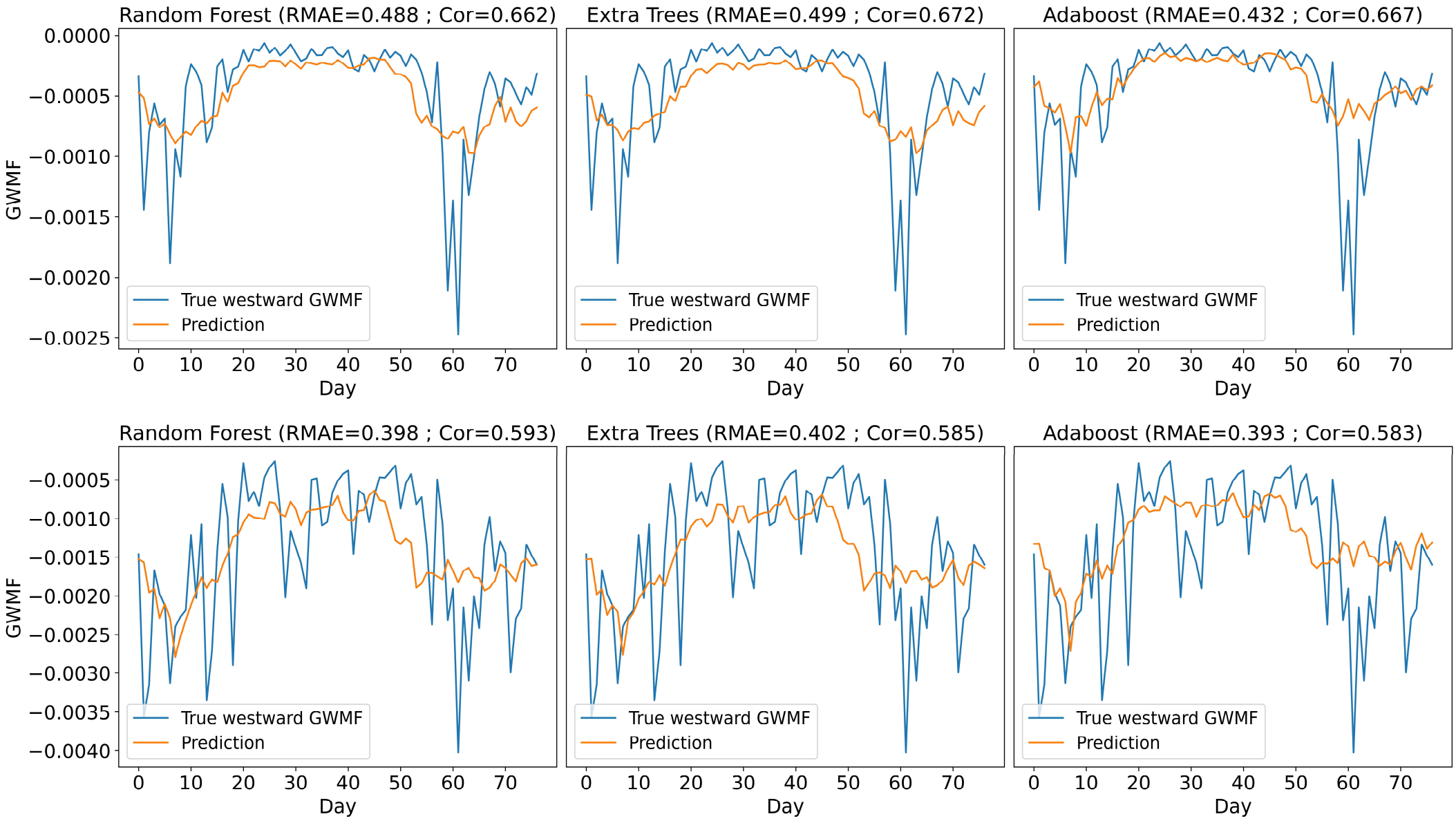
Predicted and actual absolute GWMF of HF (top) and WF (bottom) waves in 24h resolution

Eastward GWMF : Balloon 7



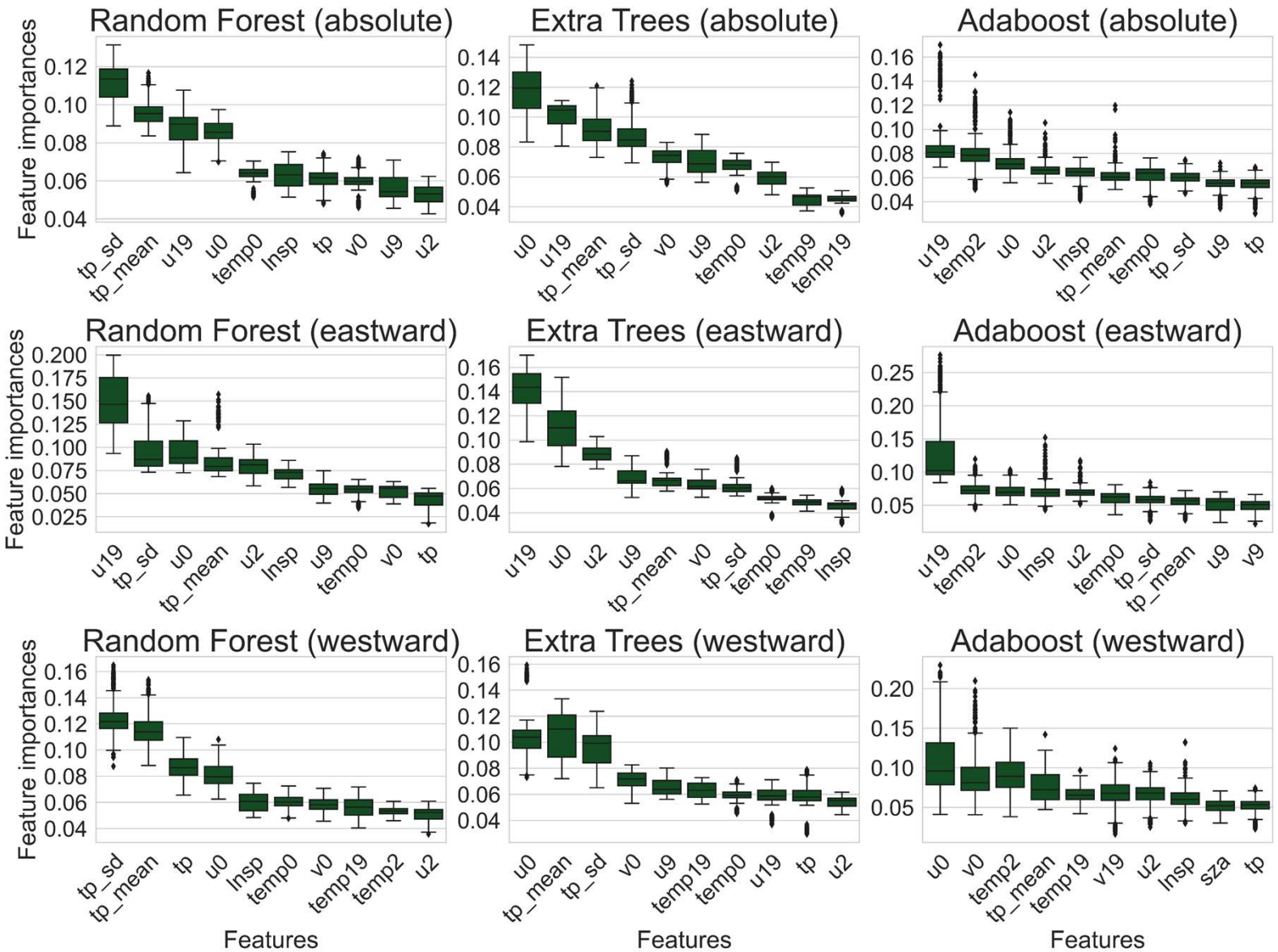
Predicted and actual eastward GWMF of HF (top) and WF (bottom) waves in 24h resolution

Westward GWMF : Balloon 8



Predicted and actual westward GWMF of HF (top) and WF (bottom) waves in 24h resolution

Feature importance : HF

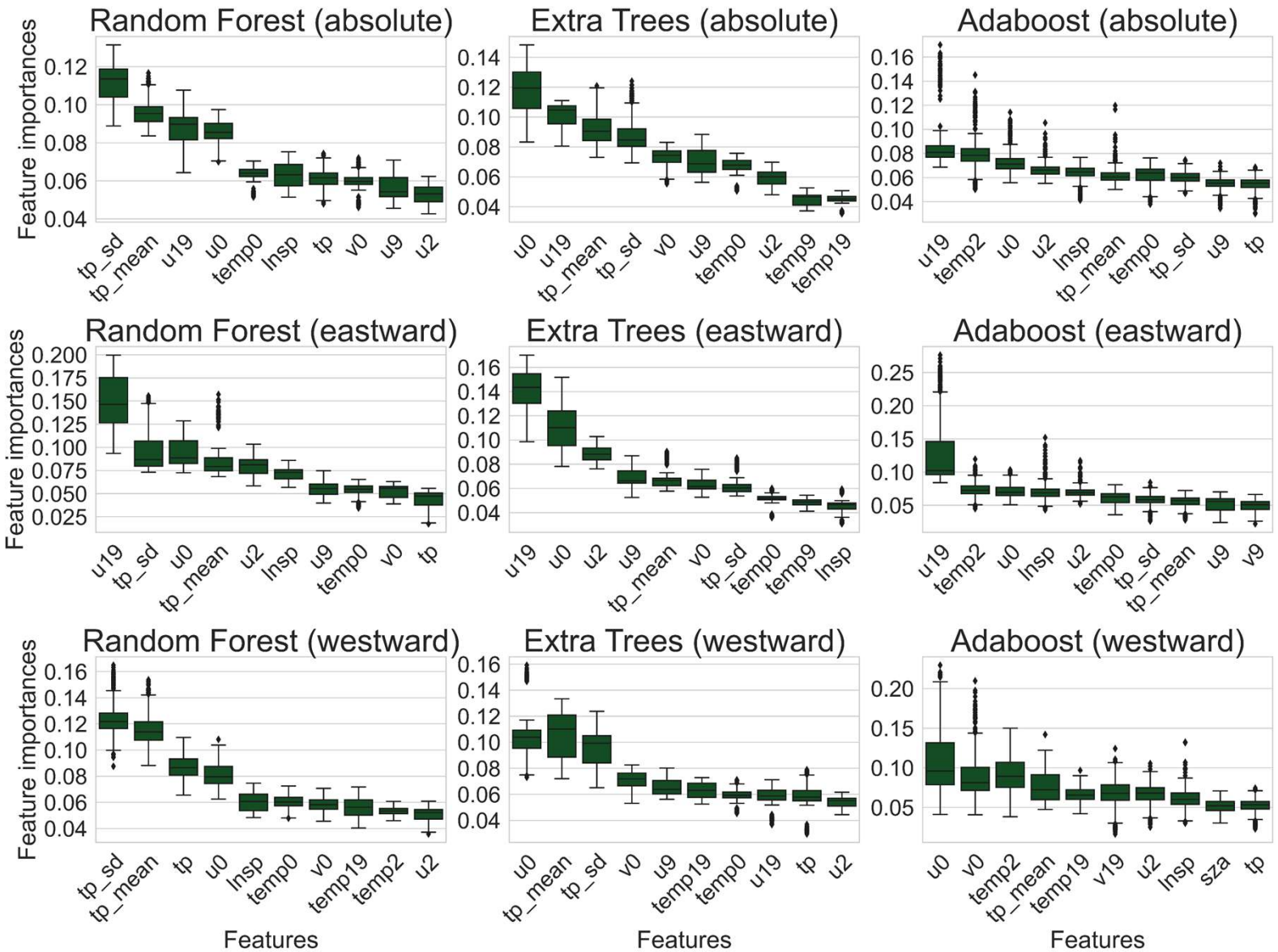


In general, precipitation and zonal wind are the most important features

Wind at balloon level u19 first in eastward case for all models

Surface wind also very informative in many cases

Feature importance : WF

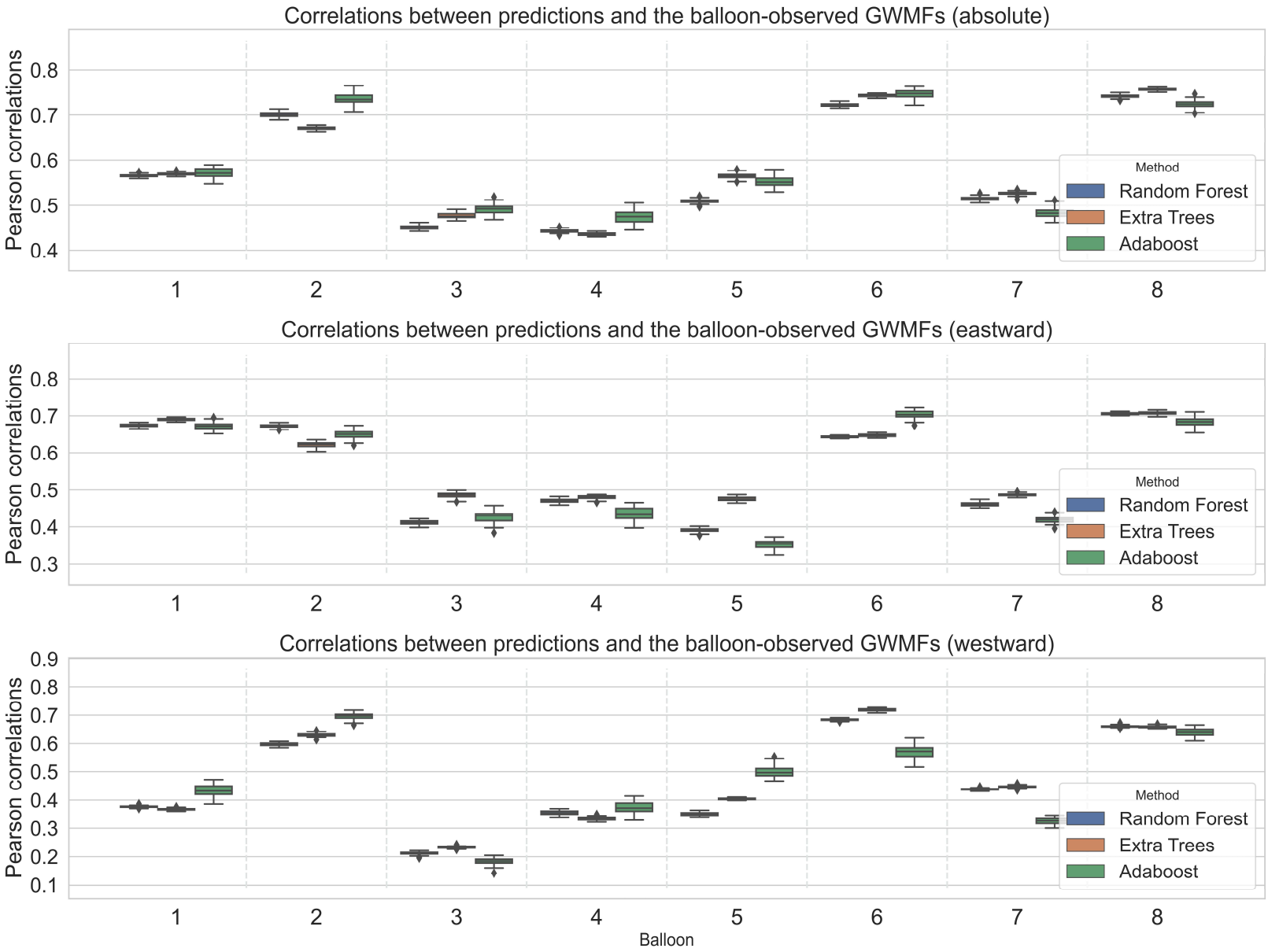


Importance of zonal wind for absolute GWMF

Wind at balloon level u19 in eastward case for all models

Precipitations more informative in westward cases

Correlations HF (50 runs)

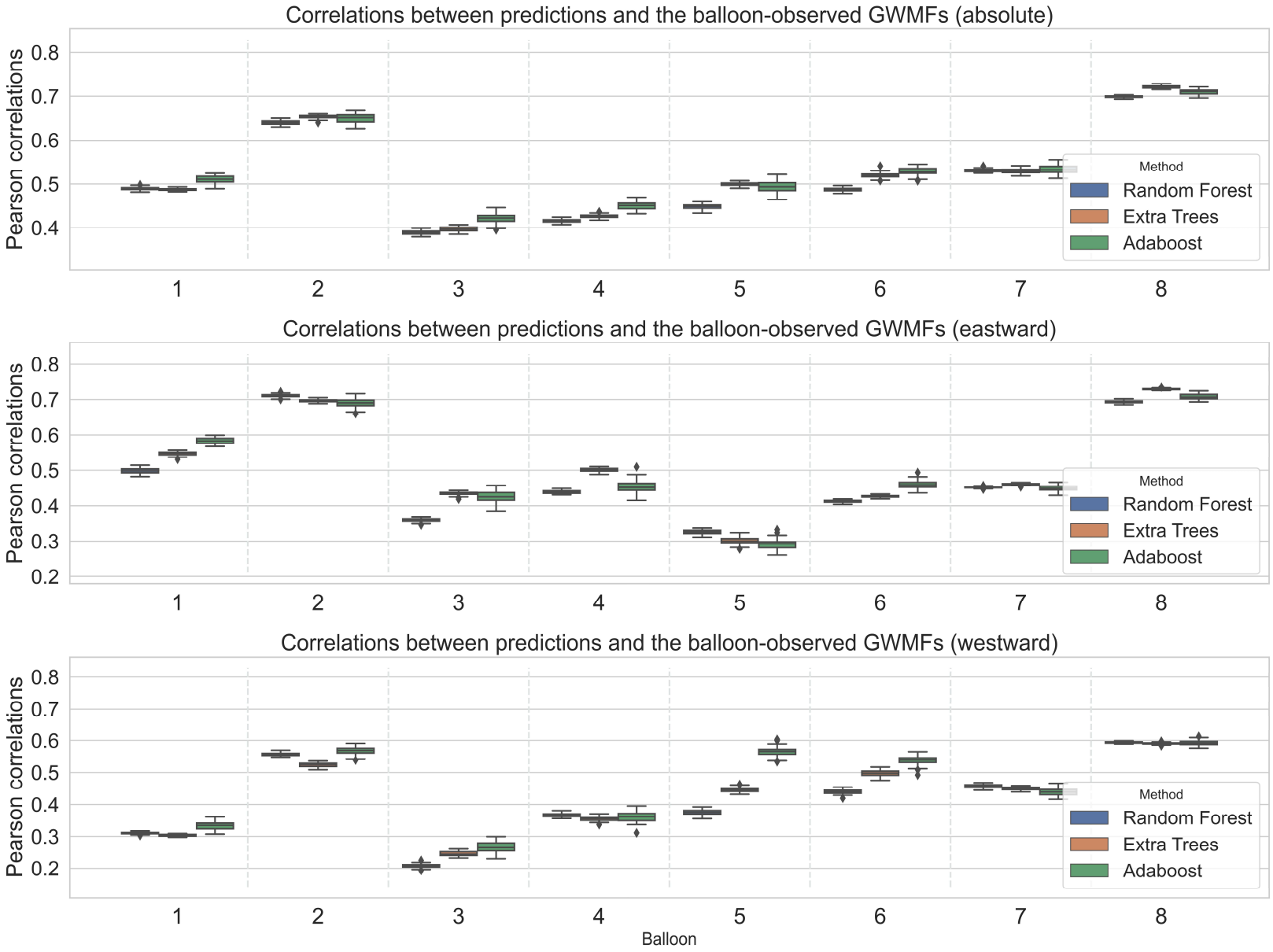


ML methods perform similarly

Balloons 2, 6, 8 well predicted (cor > 0.7)

Westward GWMF more challenging

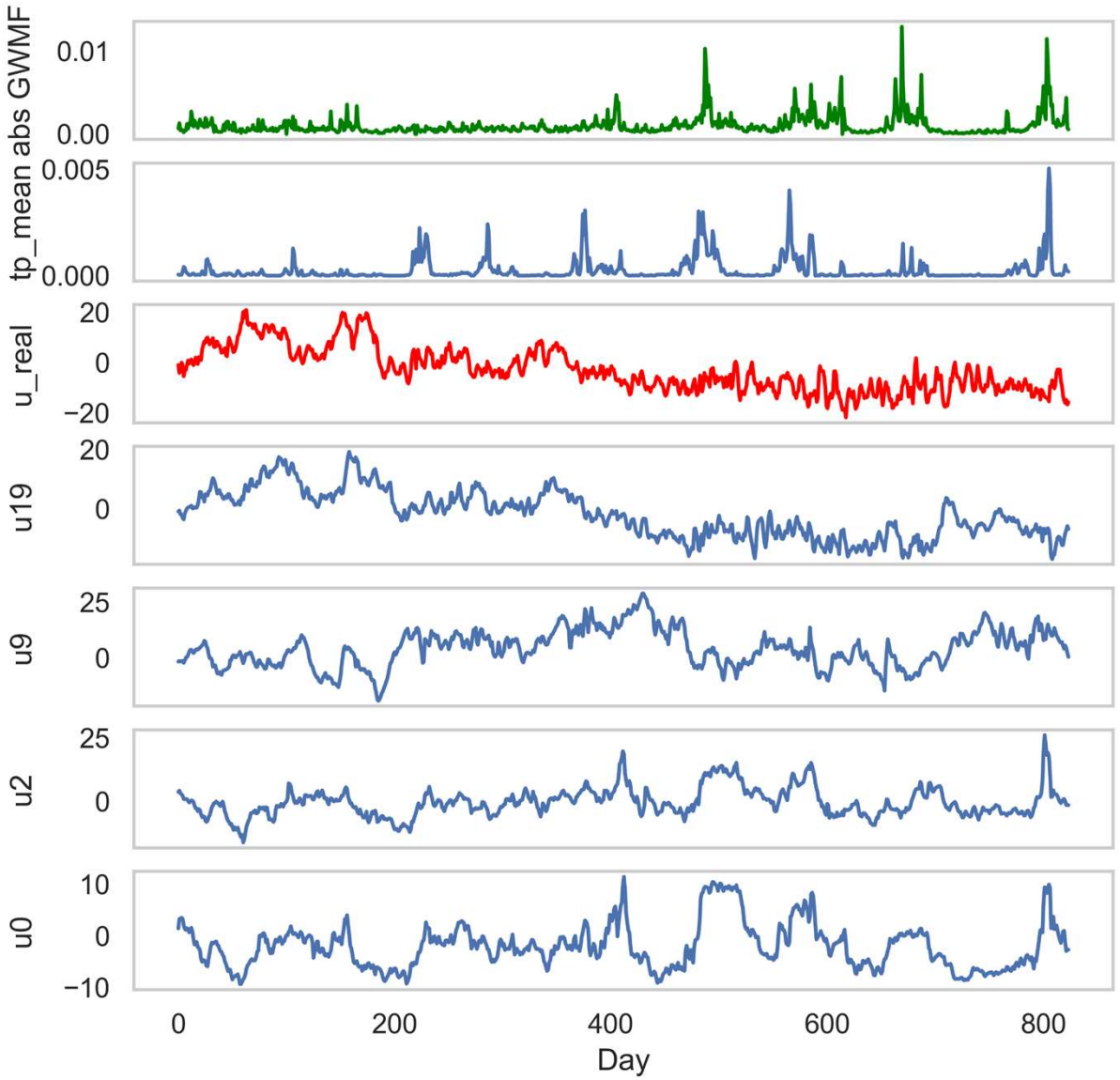
Correlations WF (50 runs)



Balloons 2 and 8 still well predicted (cor > 0.7), but not 6

Performance on WF often lower

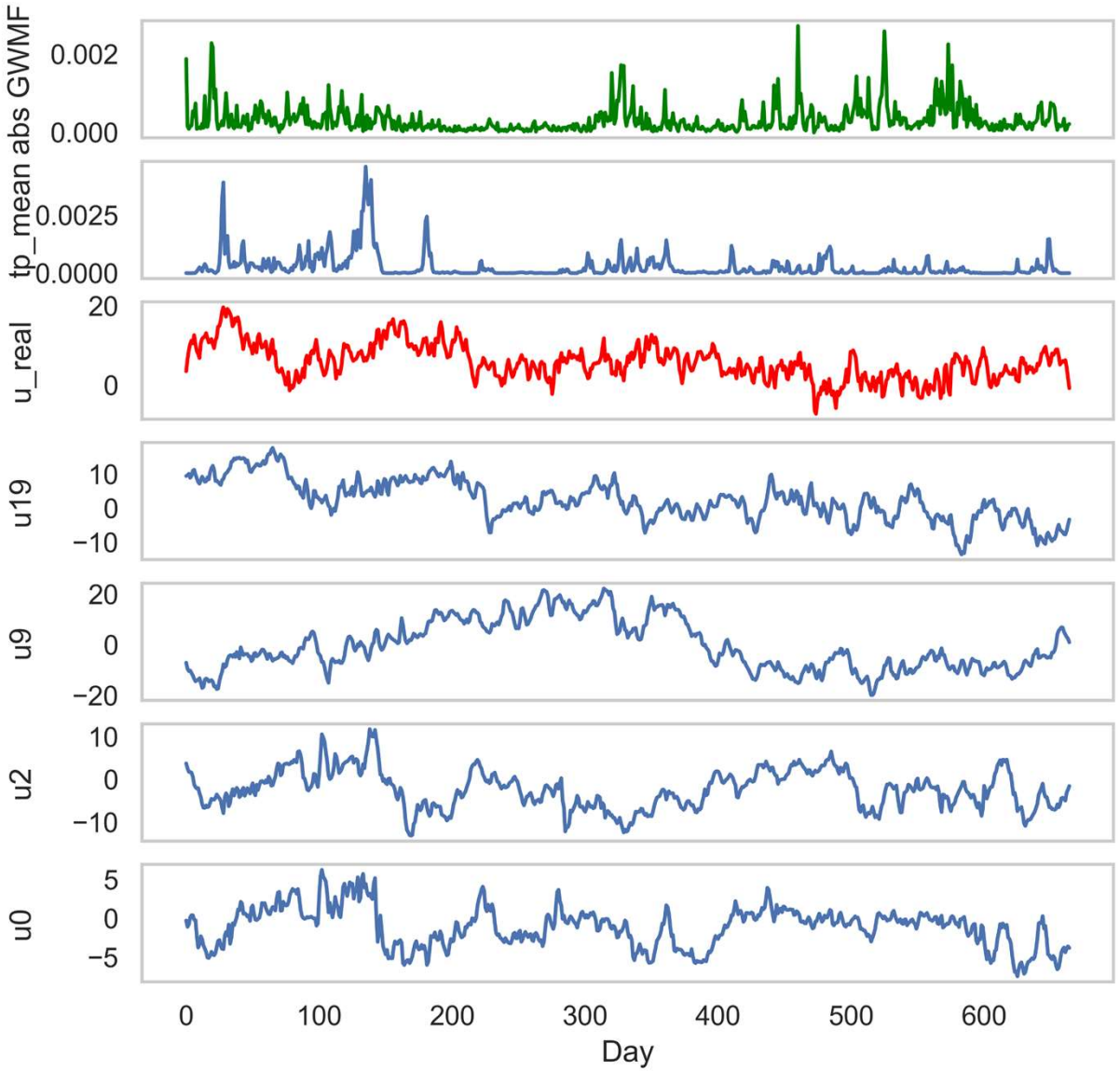
Absolute GWMF vs important variables : Balloon 2



Precipitations correspond well to GWMF

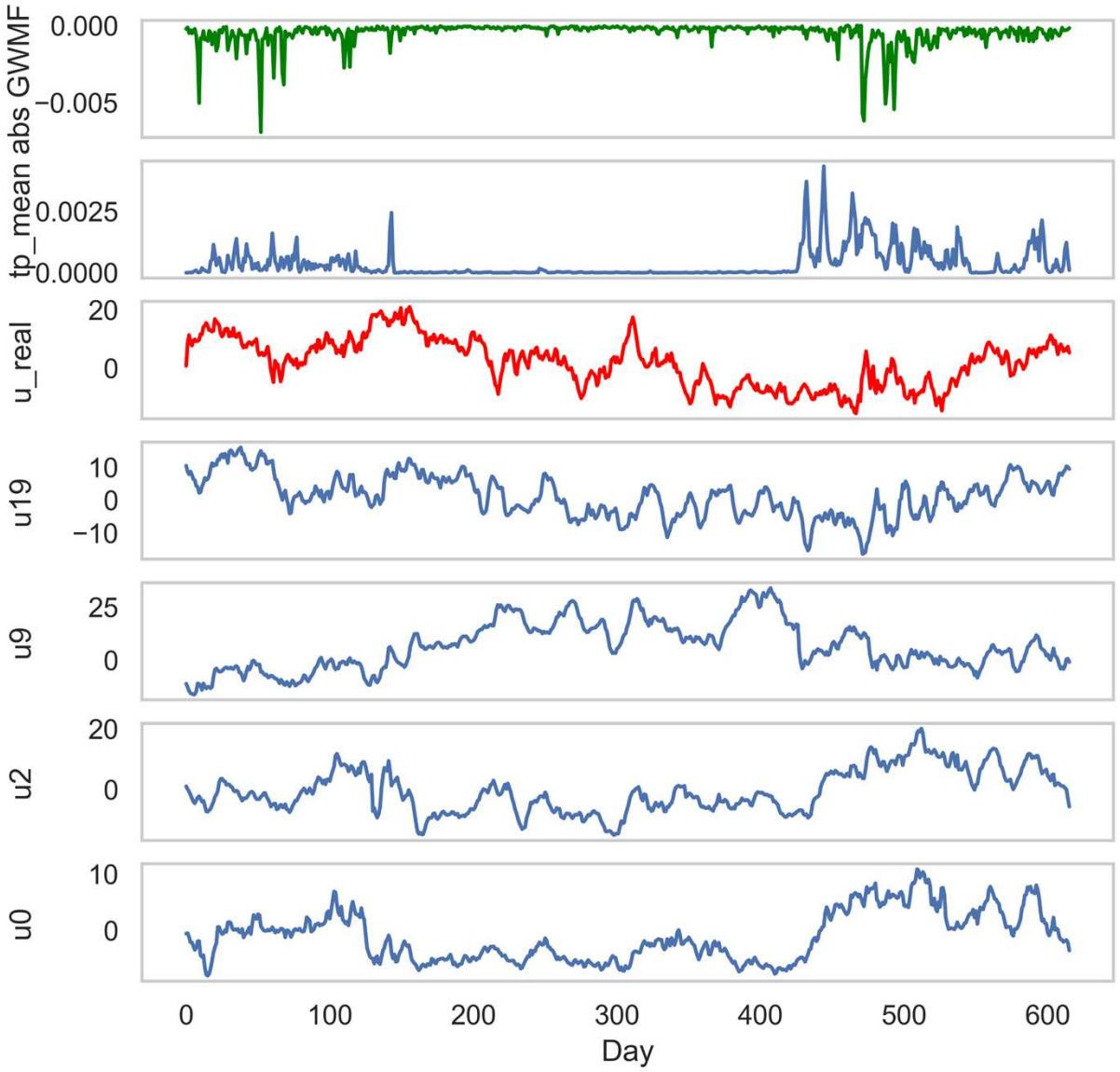
Winds seem informative as well, both at balloon level and below

Eastward GWMF vs important variables : Balloon 7



Precipitations not very informative.

Westward GWMF vs important variables : Balloon 8



Precipitations and wind seem more informative than in previous case

Remark

- Differences HF / WF ?

- Frequency determined by the angle of the phase lines :
- HF : almost vertical (gravity effective as a restoring force)
- LF : oblique, almost horizontal.

- Air motion parallel to phase lines.

- Local information corresponds well to HF waves propagating vertically.
- WF background noise difficult to link to a source.

Some conclusions

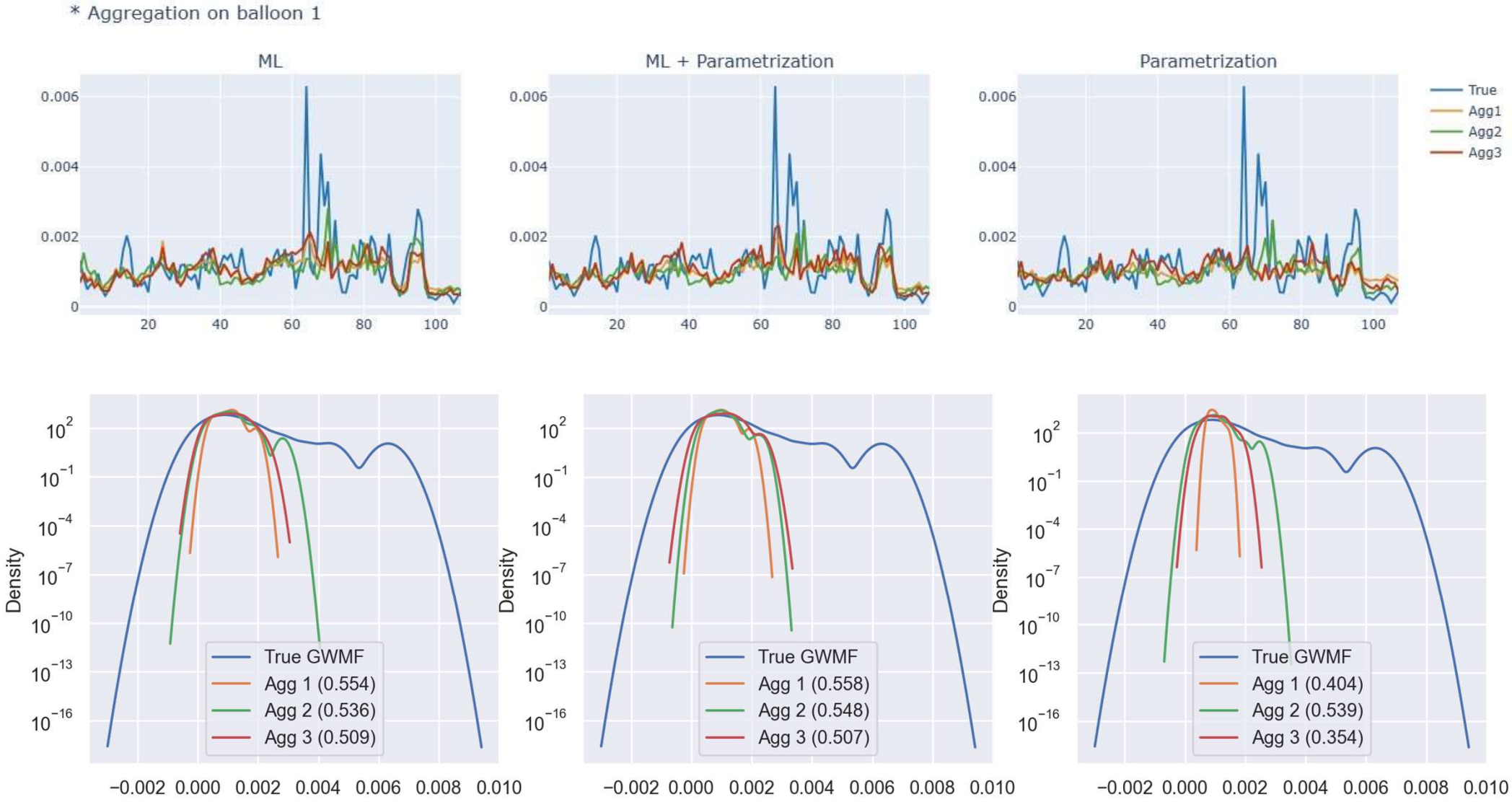
- Reconstruction of GWMF up to an encouraging level (correlation > 0.7)

→ lower bound on how much can be reconstructed from large-scale flow described by reanalysis

- Most informative variables : precipitations + zonal wind at and below balloon level
- Ocean / land

Aggregation : statistics & parameterizations

Aggregation ML & parameterization (learning on 2021 observations : Balloon 1



A few thoughts on the results

- Parameterizations catch relevant, valuable and nontrivial physics (for some balloons, some parameterizations score better)
- There is a wide variety of parameterizations.
- Machine learning extracts information from the background flow : good results, but limitations. Purely data-driven parameterizations ?

Some perspectives

More data ?
Exploration thanks to high resolution simulations.

More informative inputs ?
Different kinds of satellite image data : knowledge about convection, precipitations.