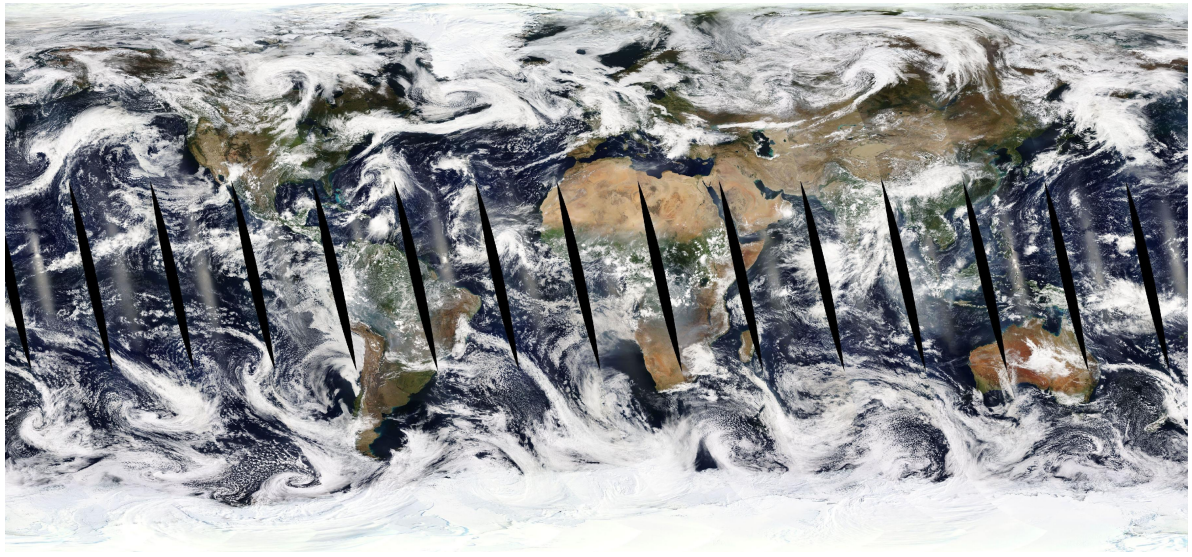
Cumulo: A Dataset for Learning Cloud Classes

Valentina Zantedeschi, Fabrizio Falasca, Alyson Douglas, Richard Strange, Matt J. Kusner, Duncan Watson-Parris

Tackling Climate Change with Machine Learning, 14/12/2019 Vancouver

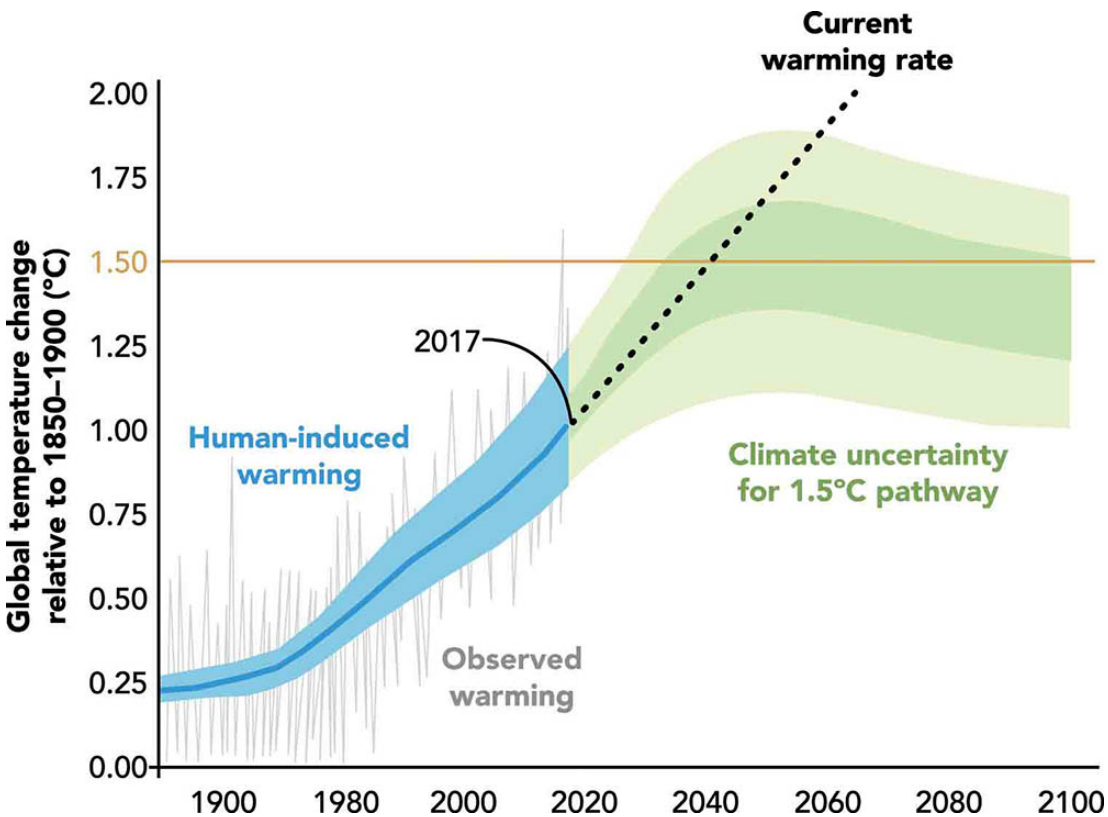


ESA Frontier Development Lab 2019 - Atmospheric Phenomena and Climate Variability challenge

Classify Clouds for Climate Projections

motivation

clouds: the greatest source of global warming uncertainty



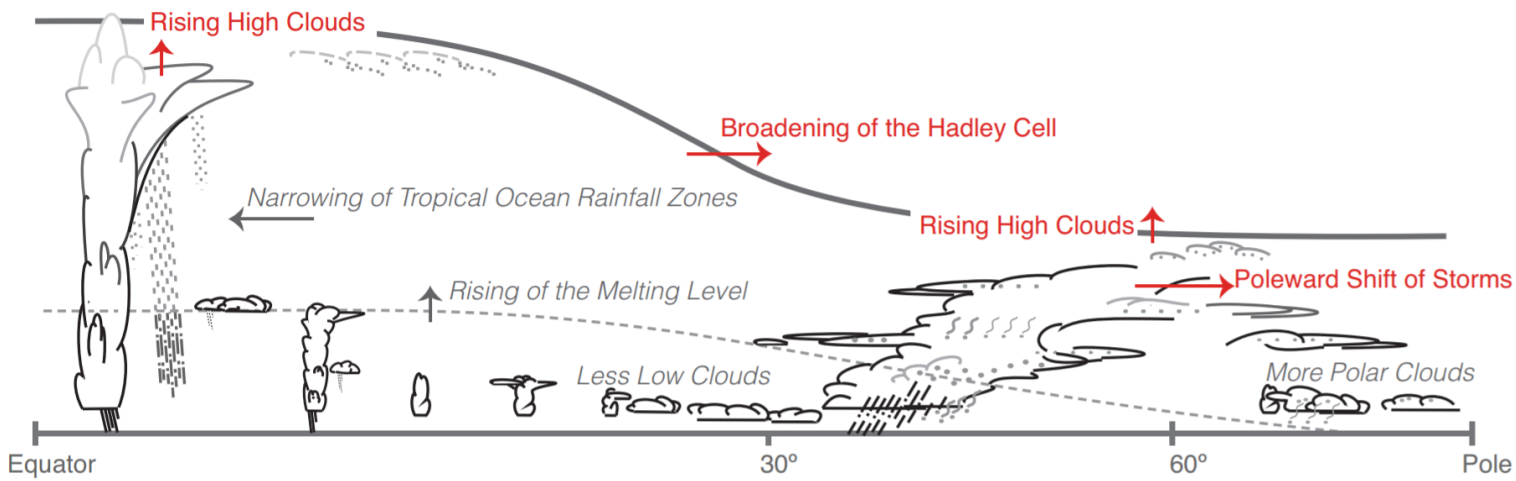
- precipitation
- radiative budget
- feedbacks

Classify Clouds for Climate Projections

motivation

To reduce uncertainty

(a) classify clouds into types and (b) study their spatiotemporal variability



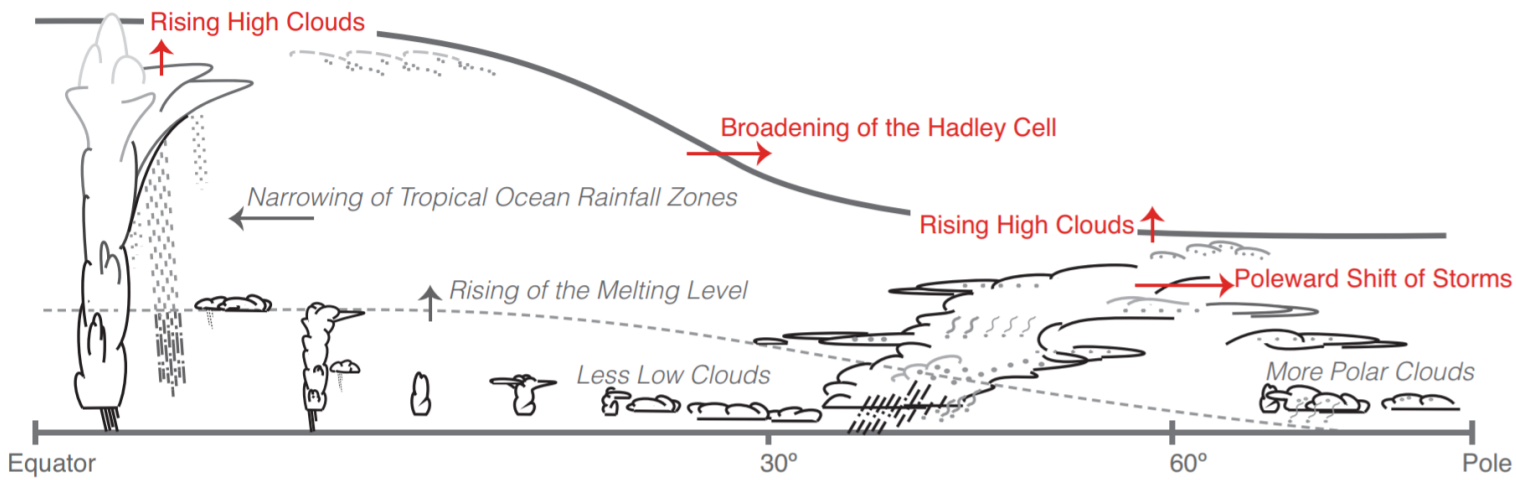
Classify Clouds for Climate Projections

motivation

To reduce uncertainty

Machine Learning — Climate Science

(a) classify clouds into types and (b) study their spatiotemporal variability



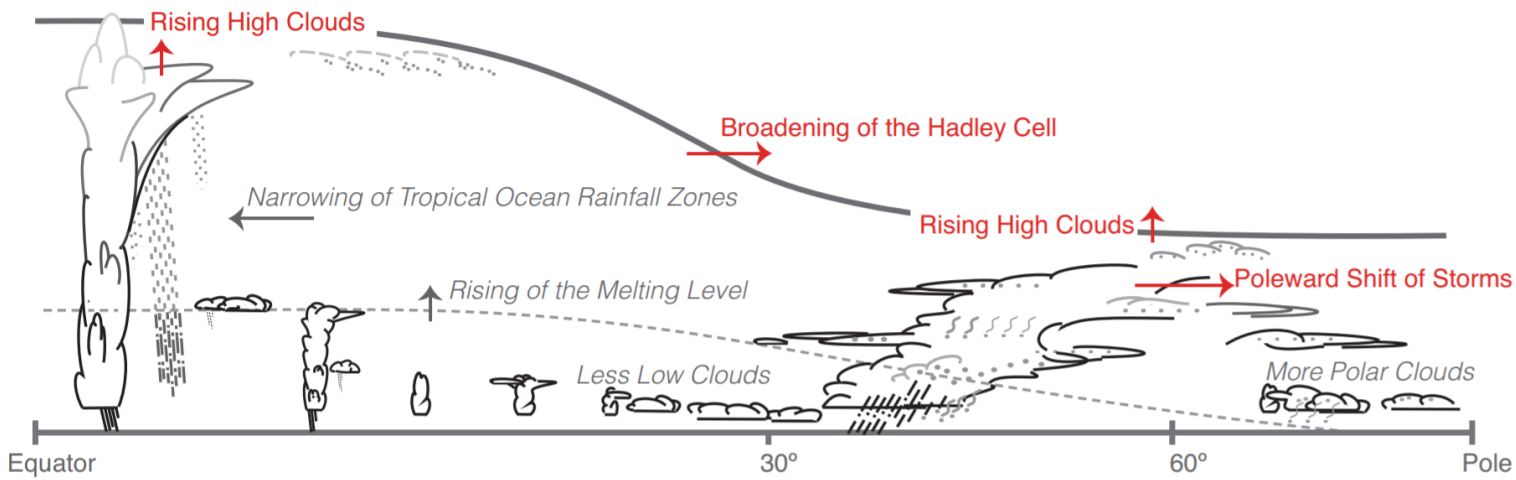
Classify Clouds for Climate Projections

motivation

To reduce uncertainty

Machine Learning — Climate Science

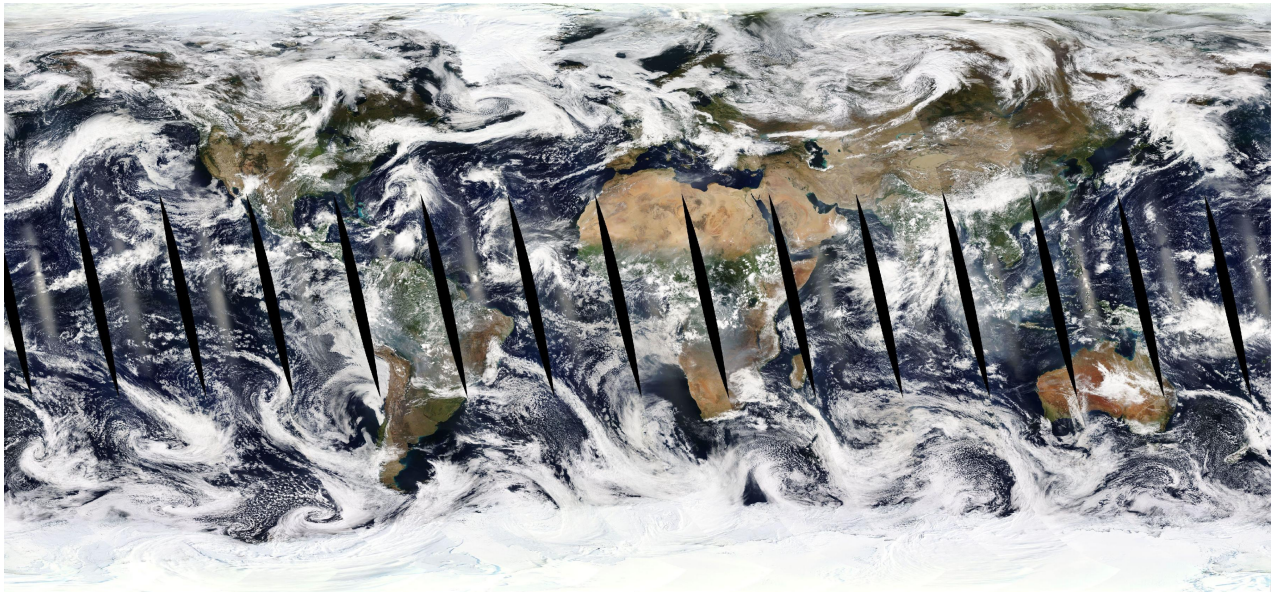
(a) classify clouds into types and (b) study their spatiotemporal variability

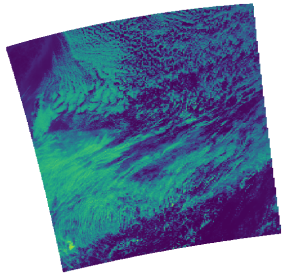


This work: introduce a new cloud classification dataset to the machine learning community to address (a)

Cumulo Dataset

features





Visible Channel

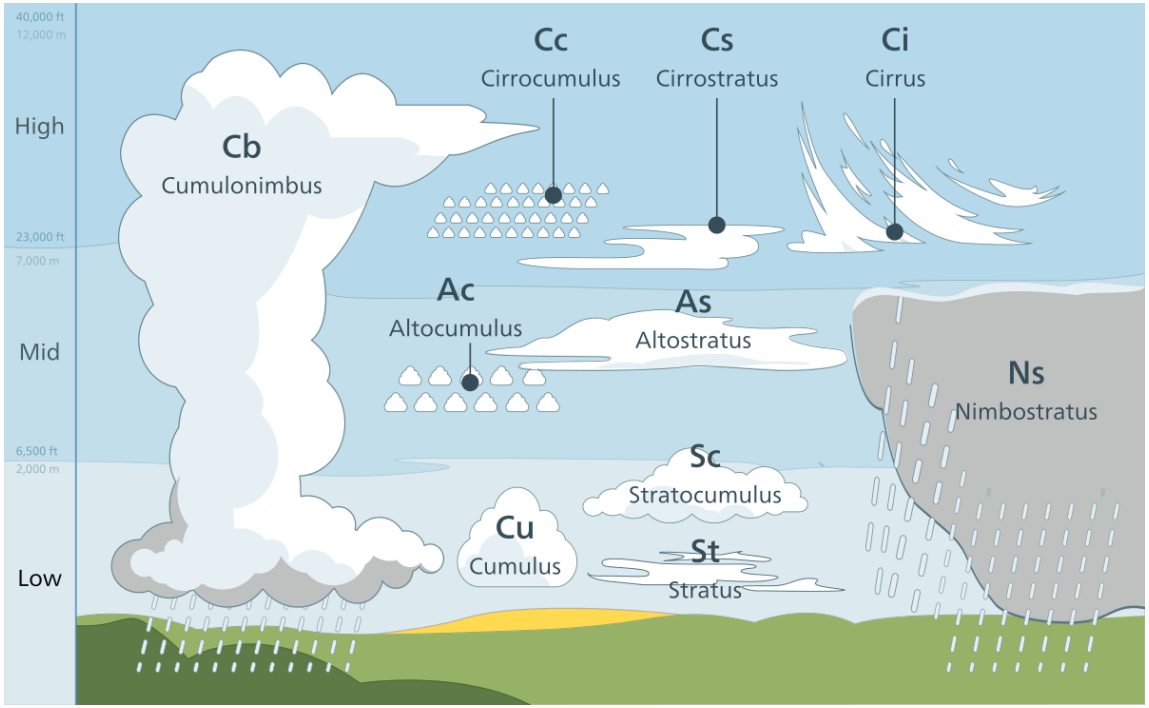
Cloud Mask

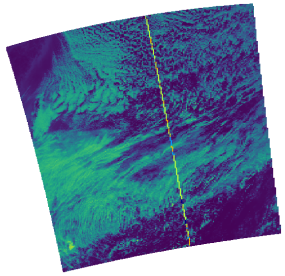
Hyperspectral Satellite Images (MODIS Aqua)

- 105,120 images of 1354 $\times$ 2030 pixels
- 22 channels (13 raw radiances, 9 computed)
- longitude, latitude, cloud mask
- near global, daily coverage for 2008
- high spatial resolution (1 km $\times$ 1 km)

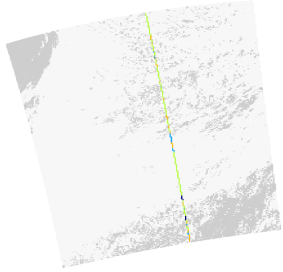
Cumulo Dataset

labels





Visible Channel



Cloud Mask

Cloud Class Tracks (CloudSat + CALIPSO)

- 8 cloud types (WMO categories: cumulus, stratus, ...)
- 1 km-wide tracks and vertical information
- daily coverage for 2008, aligned to image features

Cumulo Dataset

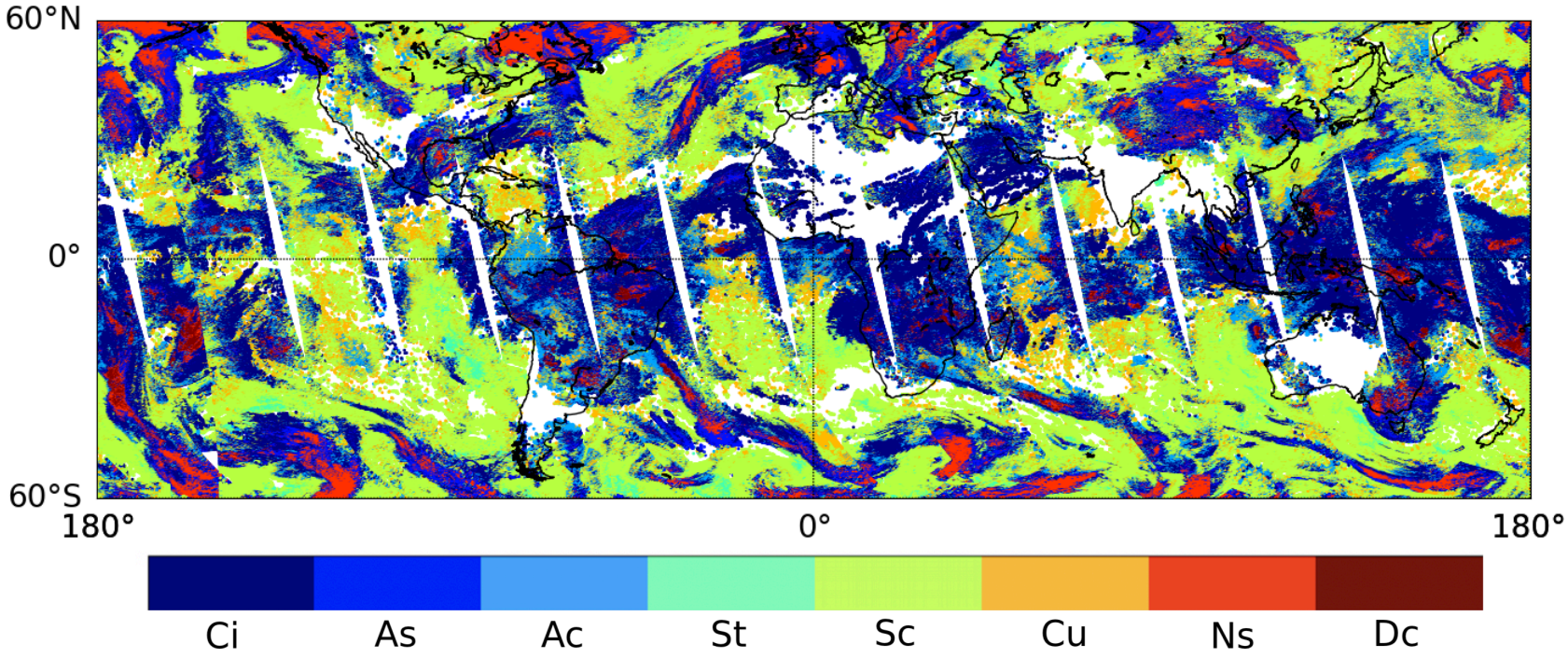
challenges

1. Supervision is available only 1 every 1354 pixels:
   weakly-labelled data
2. Pixels can be annotated with multiple types of clouds:
   multi-labelled data
3. Many cloud classes are underrepresented:
   class imbalance
4. Some channels are available only at daytime:
   missing data
5. Many cloud types have sub-types:
   unsupervised learning

Baseline Performance Analysis

standard evaluation metrics

Hybrid Invertible Residual Network [1] on 3x3 tiles



	Ci	As	Ac	St	Sc	Cu	Ns	Dc	Mean
Accuracy (%)	81.30	84.50	88.29	97.73	88.90	92.40	90.92	98.84	90.36
F1 score	0.68	0.43	0.45	0.58	0.80	0.40	0.47	0.58	0.55
IoU index	0.52	0.28	0.29	0.41	0.66	0.25	0.31	0.41	0.39

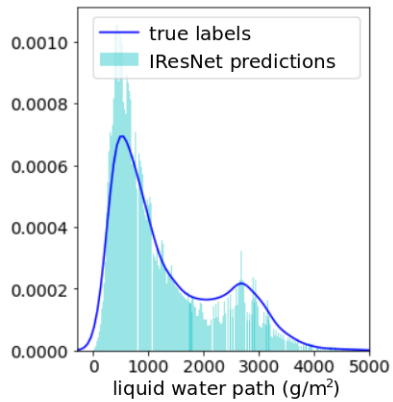
[1] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Hybrid models with deep and invertible features. *ICML*, 2019.

Baseline Performance Analysis

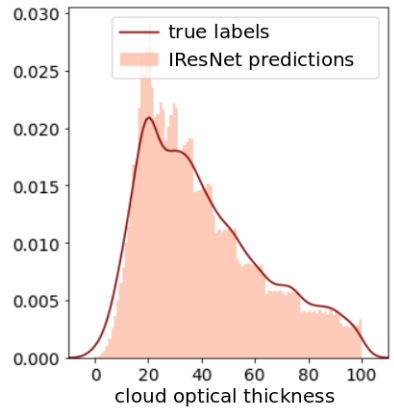
climate-based evaluation

Deep Convection clouds

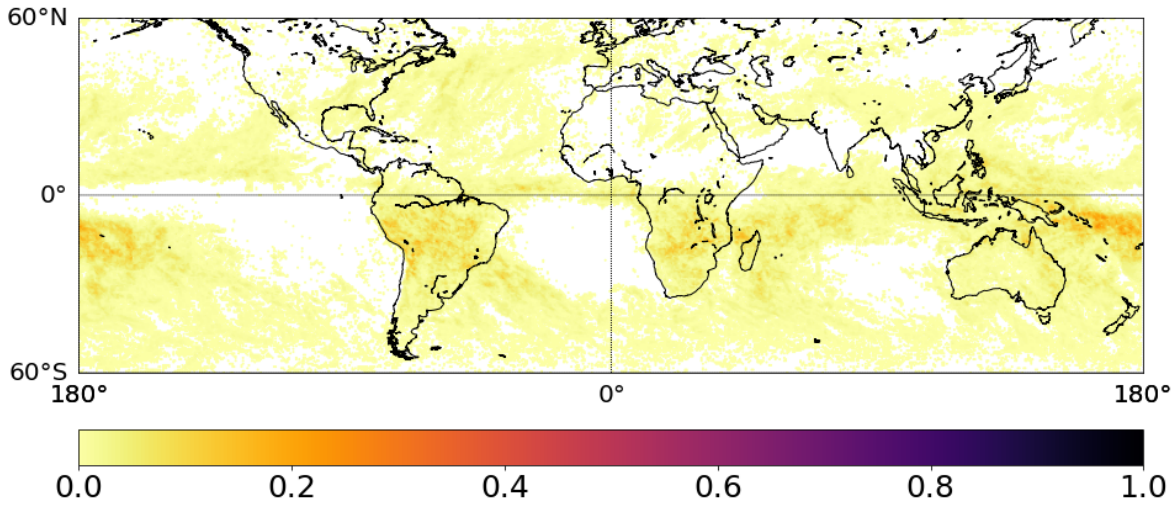
liquid water path



cloud optical thickness



$0.35 \cdot 10^{-1}$ KL divergence $0.47 \cdot 10^{-1}$



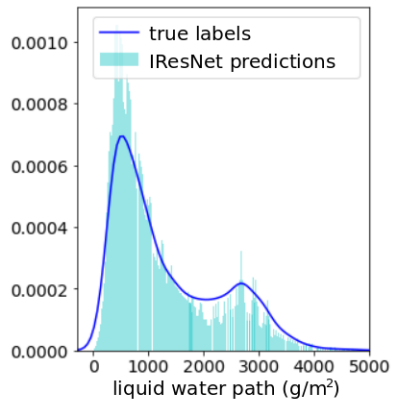
cloud occurrence

Baseline Performance Analysis

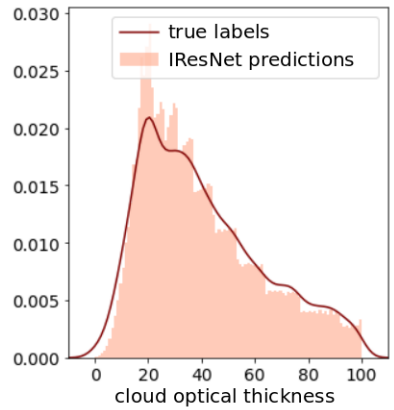
climate-based evaluation

Deep Convection clouds

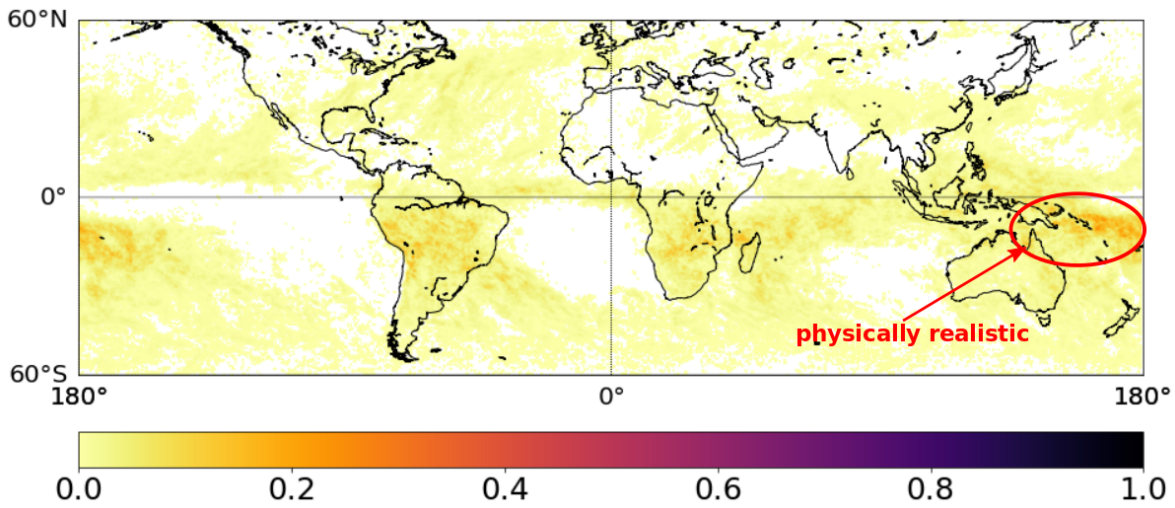
liquid water path



cloud optical thickness



$0.35 \cdot 10^{-1}$ KL divergence $0.47 \cdot 10^{-1}$



Release expected for the 30th of January

Thank you for your attention!

link to paper: `https://arxiv.org/abs/1911.04227`