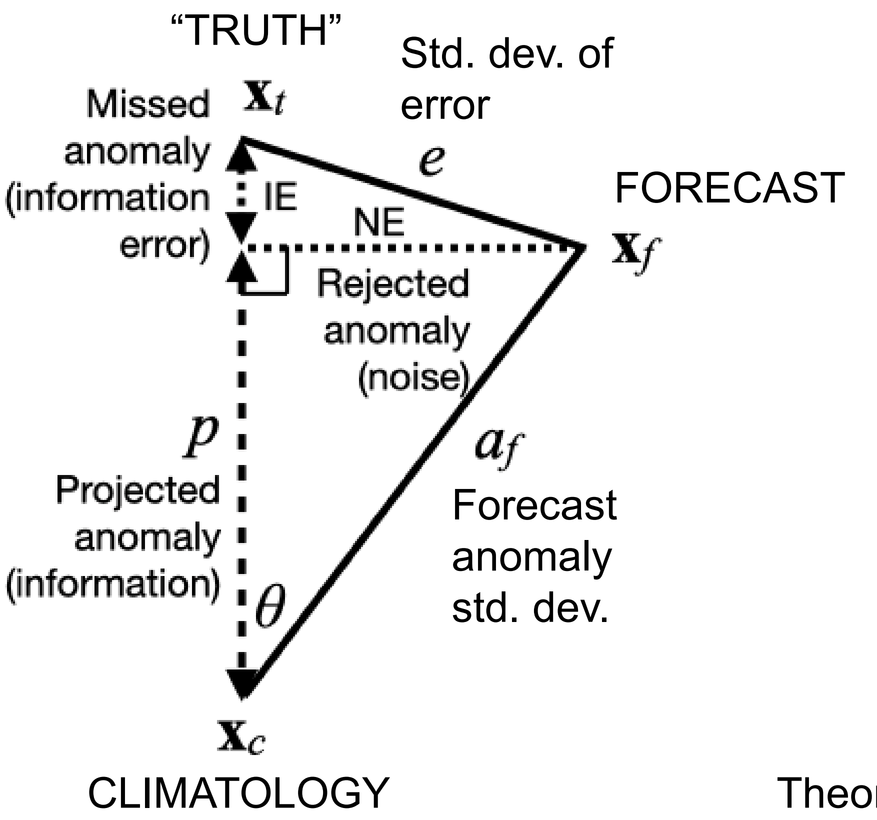
Hybrid data assimilation and machine learning

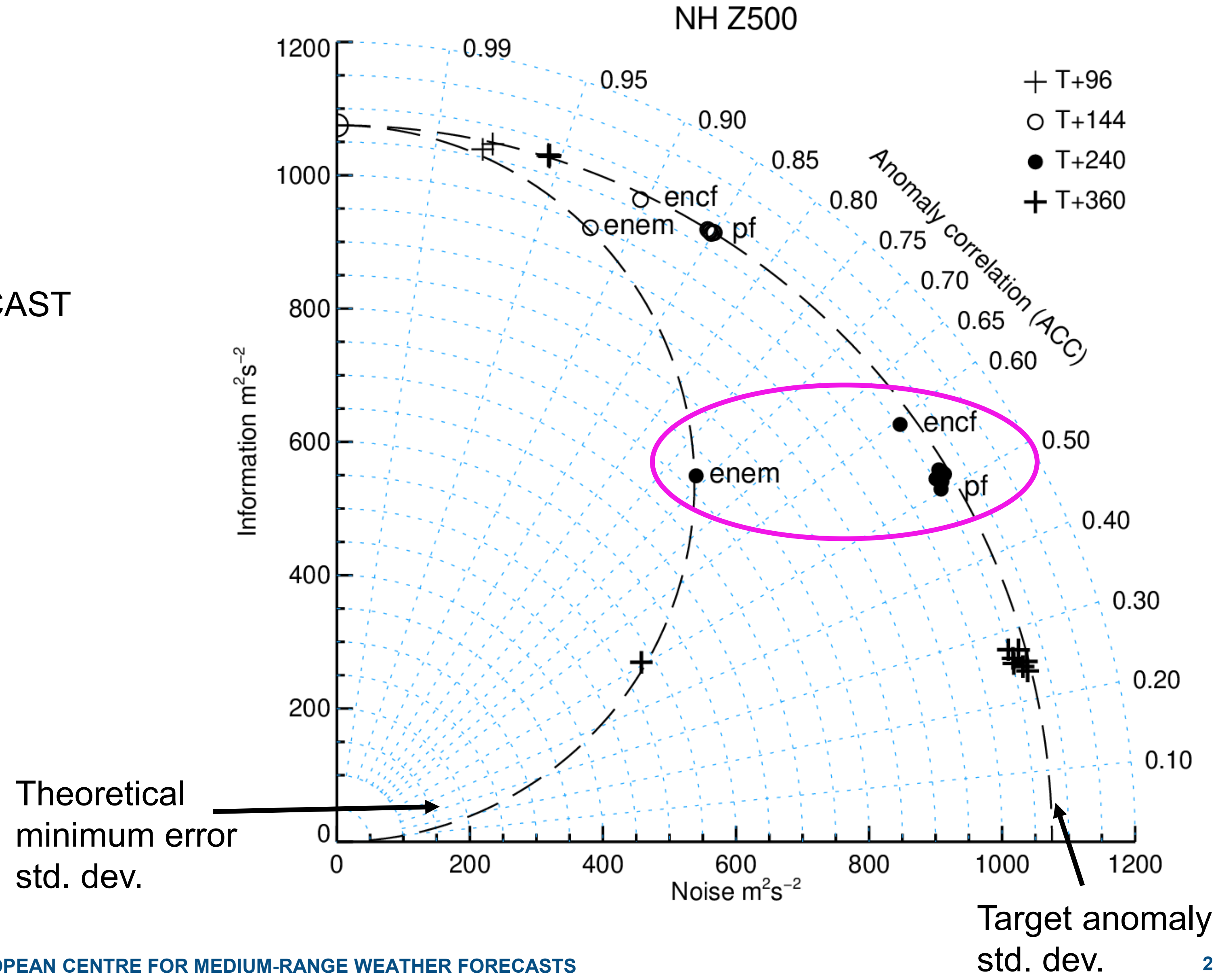
ECMWF workshop on data assimilation: initial conditions and beyond, 10th April 2025

Alan Geer, ECMWF

“Information, noise, correlation” diagram:

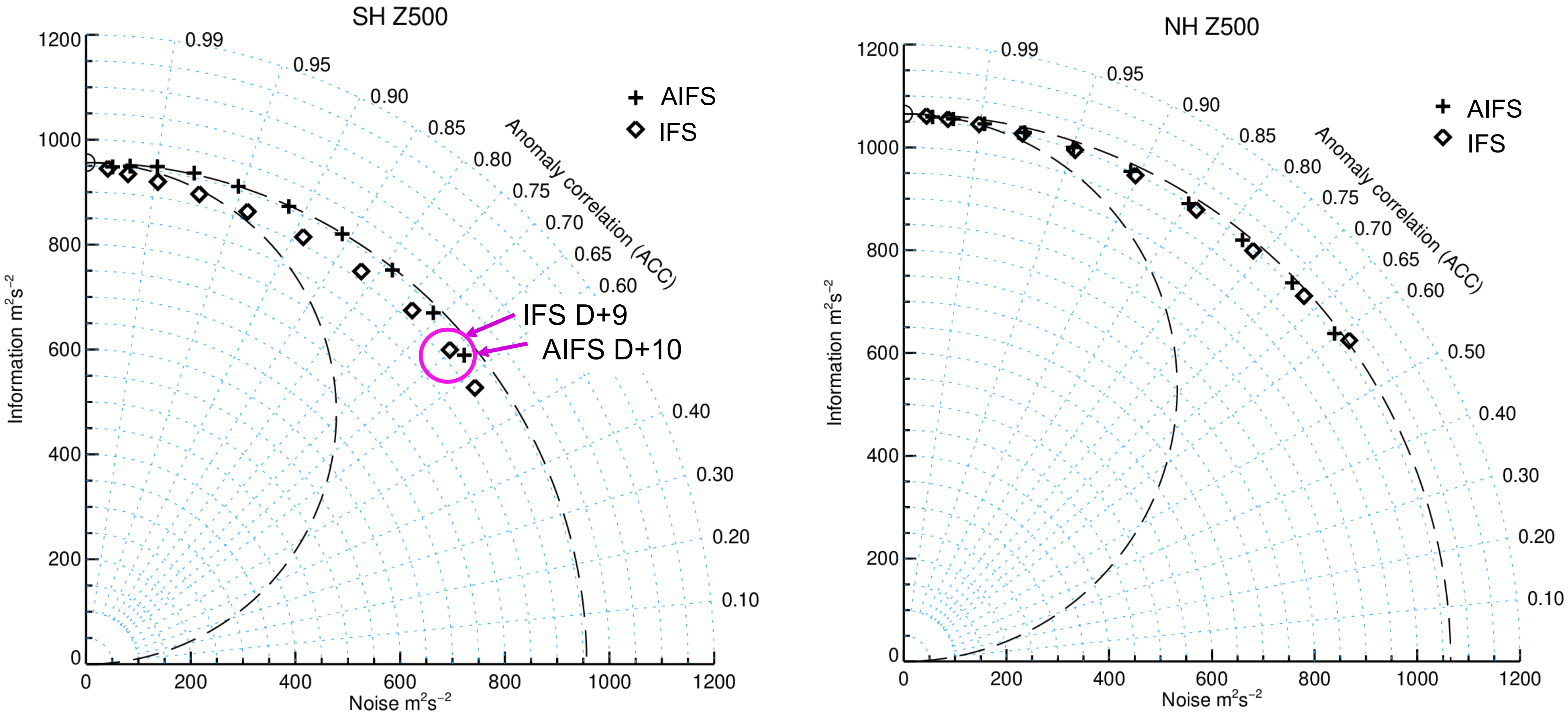
Bonavita + Geer (2025) building on FTZP24



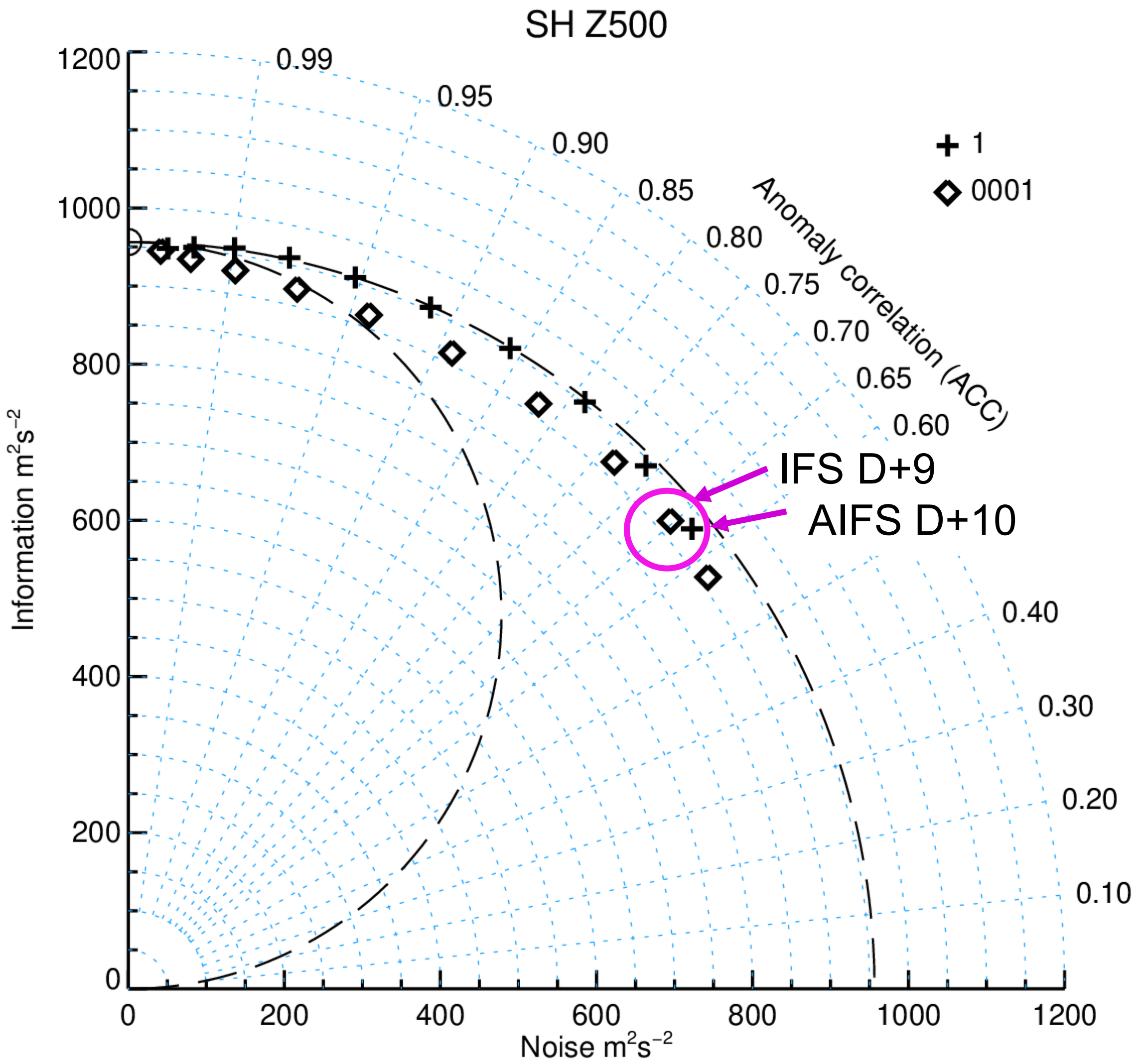


10-day AIFS forecast is close to 9-day IFS forecast in the SH summer

JFM 2025



10-day AIFS forecast is close to 9-day IFS forecast in the SH summer



- This great result is >75% due to physical data assimilation!
  - Training data is ERA5 and ECMWF operational analysis
  - Initial conditions are the ECMWF operational analysis.

- -> Medium range forecasting is possible using lower dimensionality than we thought:

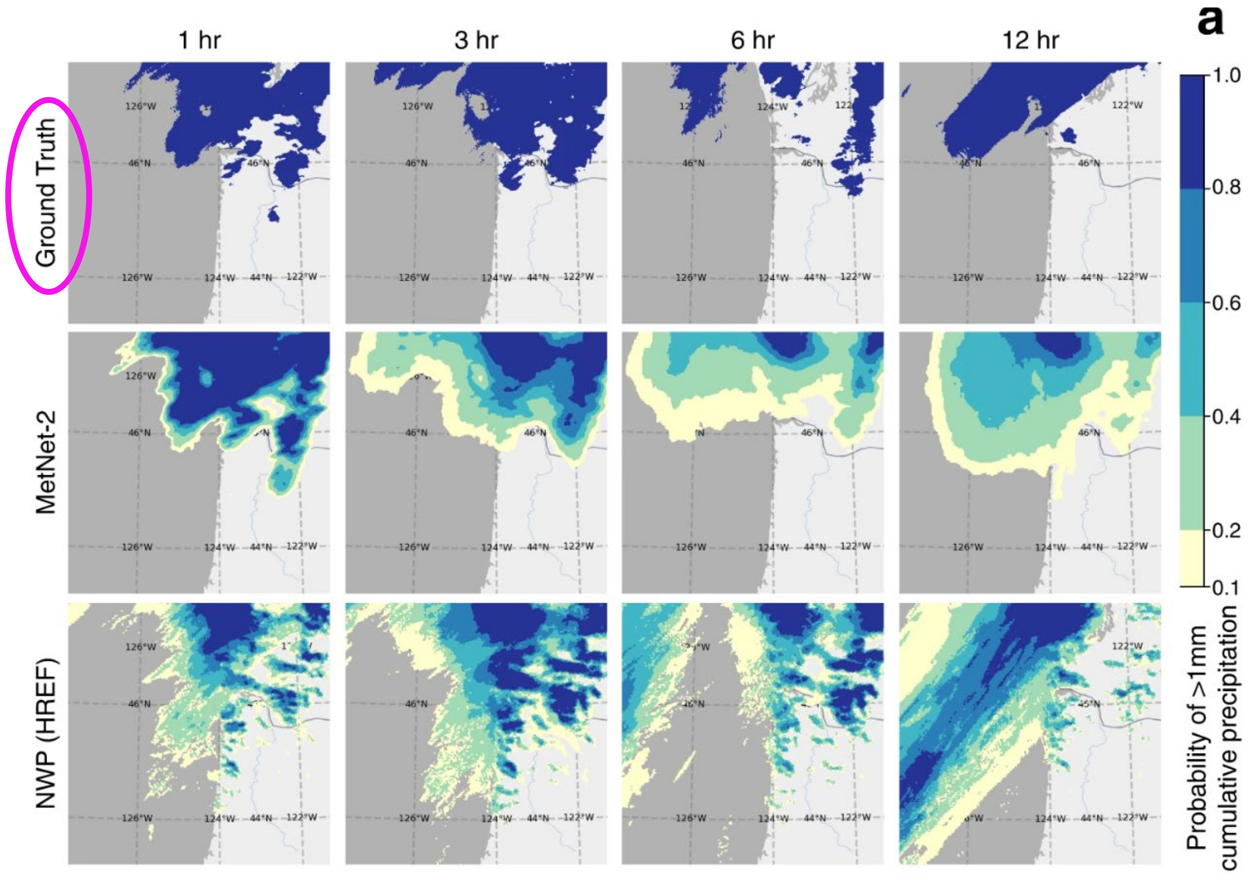
	Horizontal	Vertical	Timestep
IFS	8-9 km	137 levels	7.5 min
AIFS	36 km	13 levels	6 hour

  - Backing up older results eg https://doi.org/10.1002/qj.613
  - Machine learning creates an optimised statistical representation of the atmosphere: “latent space”

- **-> The physical model is not good enough and needs to be improved <u>using observations</u>**

Can we do without physical model or data assimilation and instead directly forecast from (and to) observations?

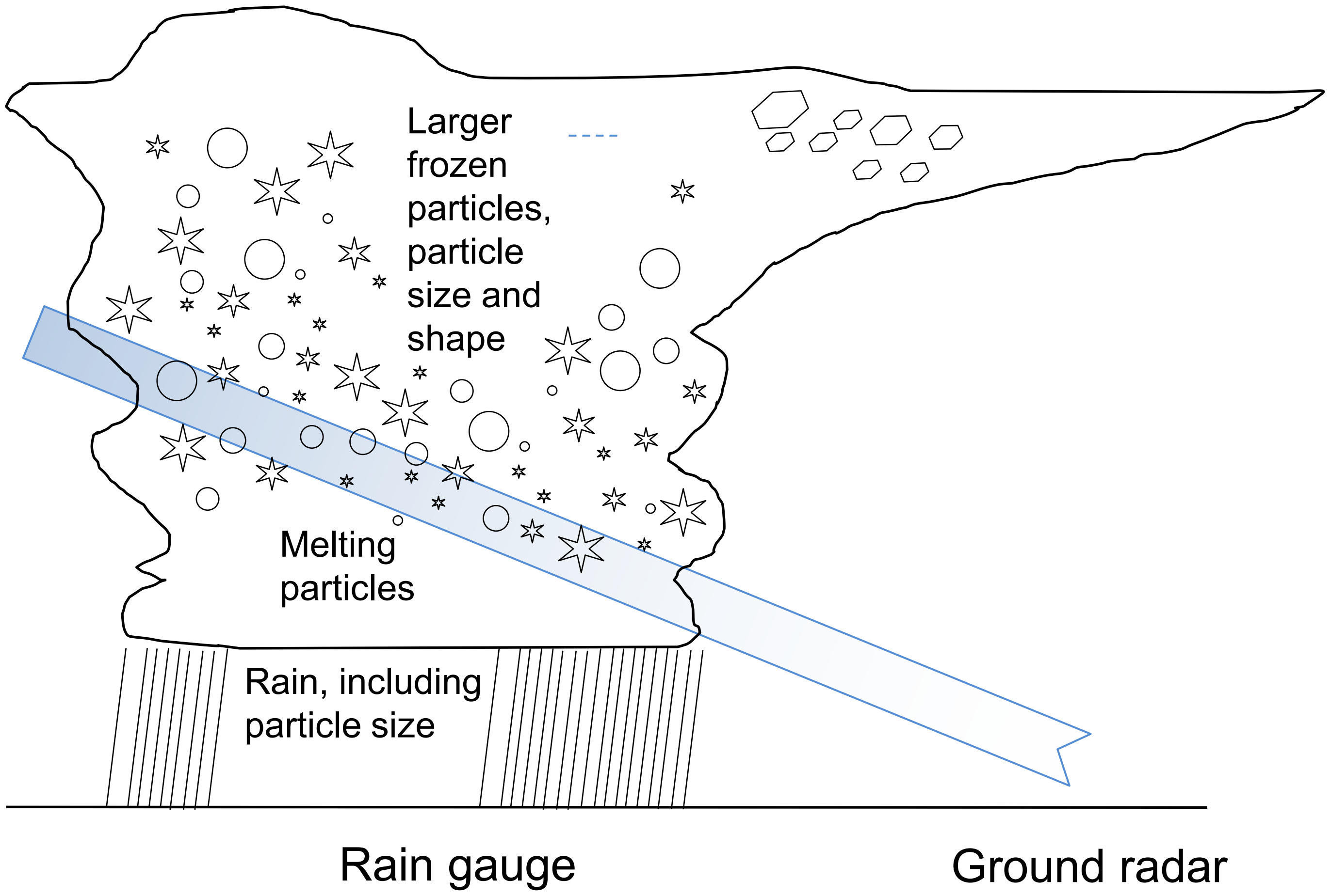
- Google MetNet-2 is trained to forecast from/to precipitation "observations" from gauge and radar (although it does use some NWP information for initial conditions)
- Aardvark Weather replaces data assimilation
  - DA emulation is trained on conventional and satellite observations with ERA5 as a target
  - Aardvark Z500 forecasts are about 1 day behind ECMWF physical forecasts
  - Allen et al., 2025, "End-to-end data-driven weather prediction", https://doi.org/10.1038/s41586-025-08897-0
- AI-DOP at ECMWF goes from observation to observation with no input from physical NWP
  - Mihai's talk or Arxiv: https://doi.org/10.48550/arXiv.2412.15687

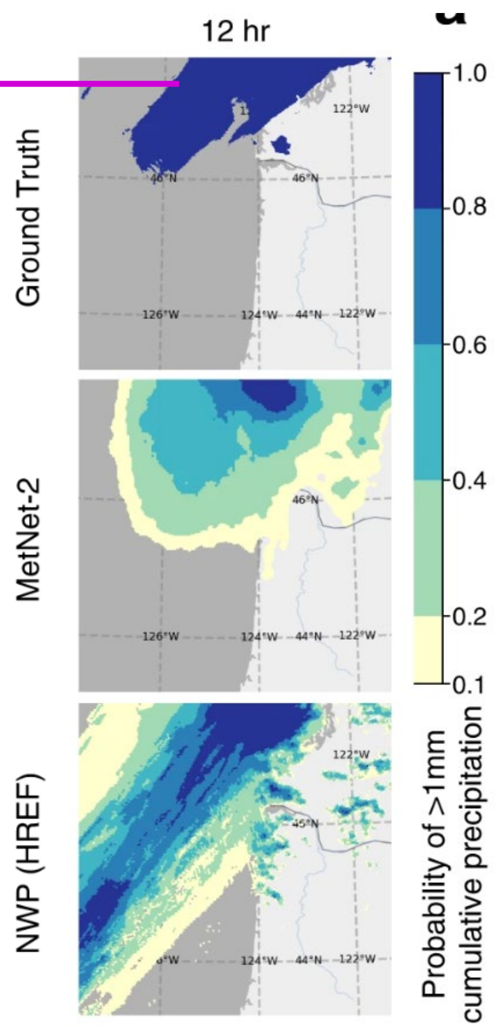


MetNet-2: CC BY 4.0 reproduction from Espeholt et al. (2022, "Deep learning for twelve hour precipitation forecasts) https://doi.org/10.1038/s41467-022-32483-x)

What actually is "ground truth"? Precipitation example / MetNet-2

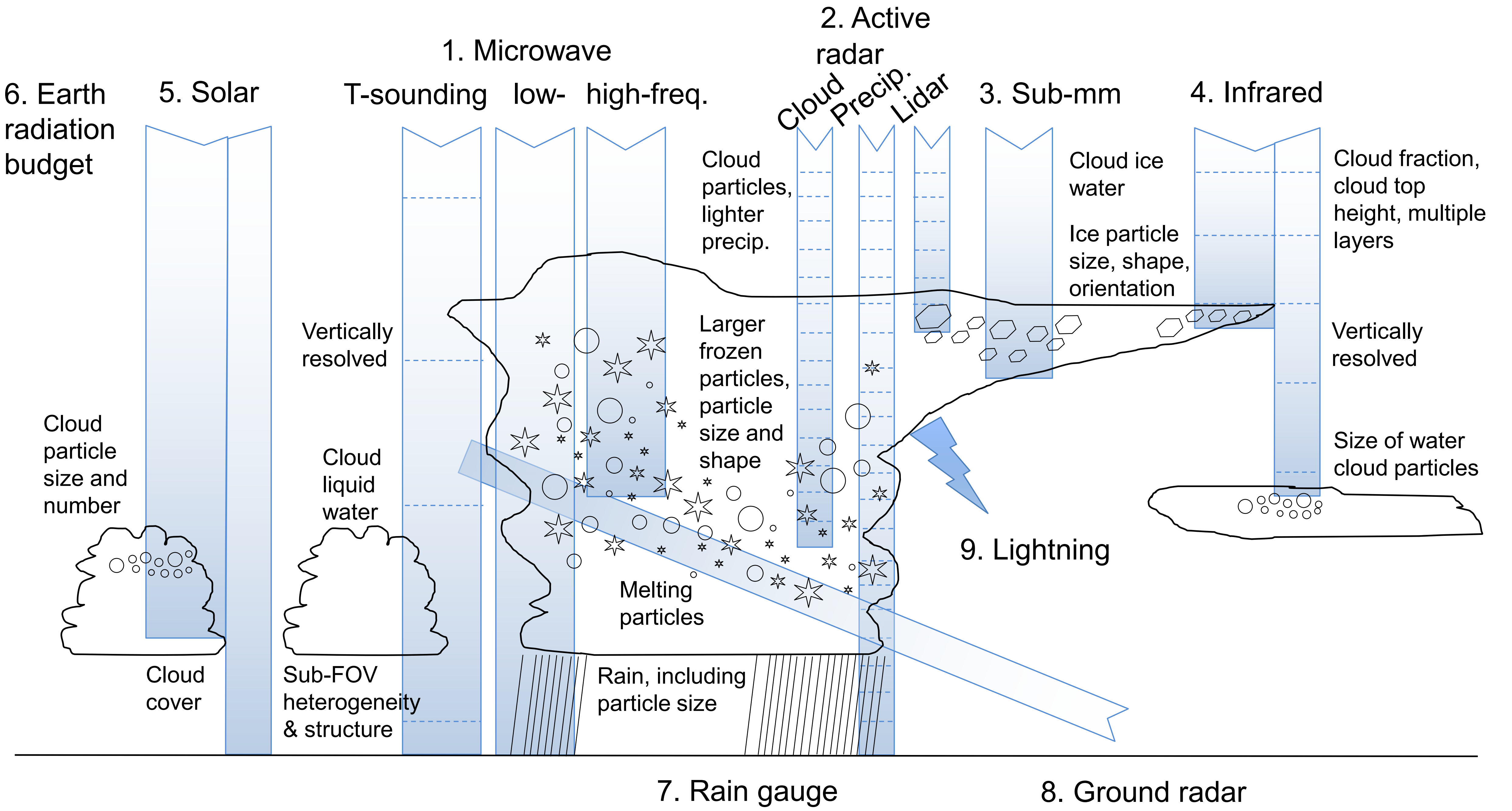
Multi-radar/multi-sensor system (MRMS).
https://www.nssl.noaa.gov/projects/mrms/



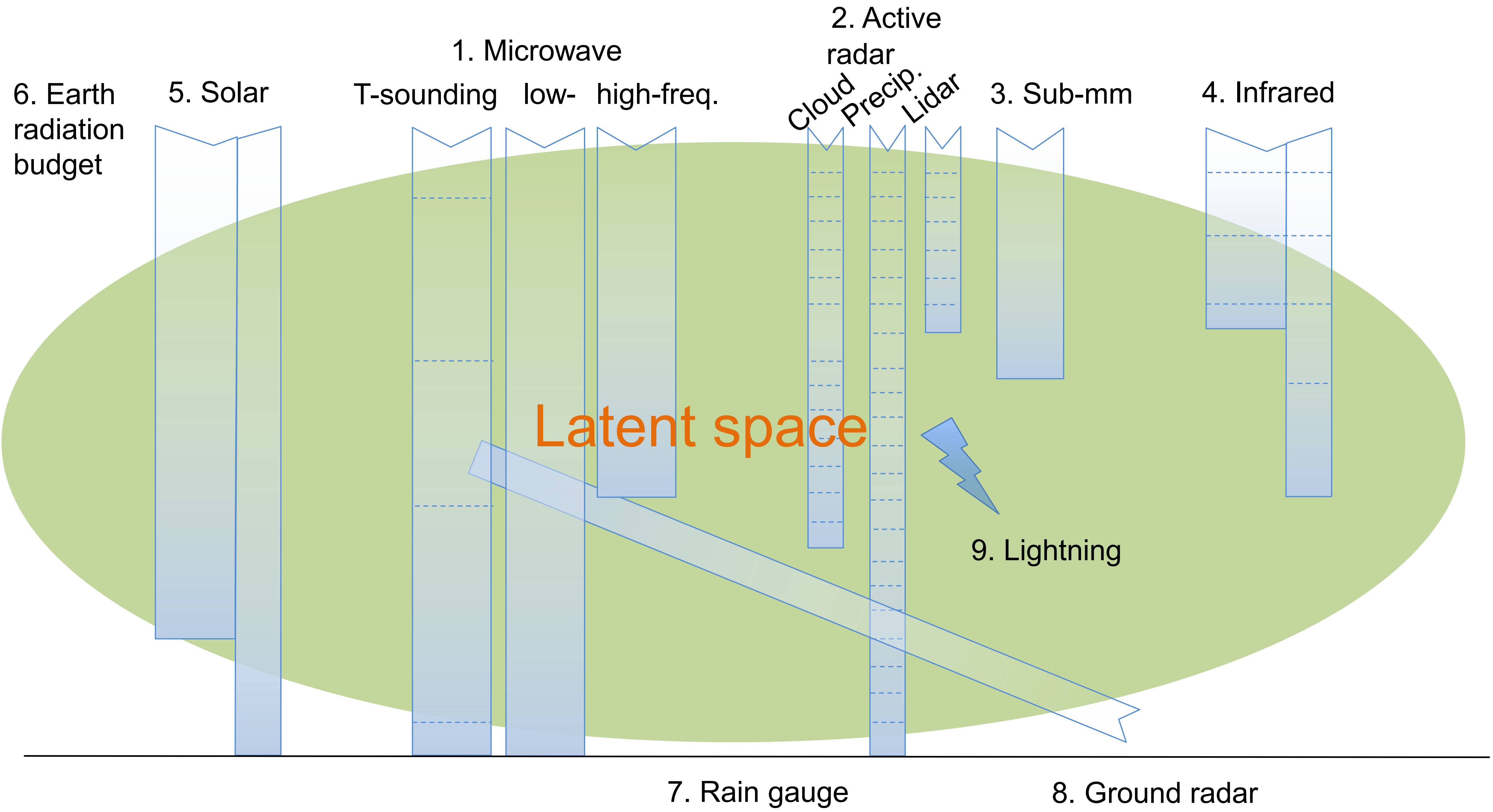


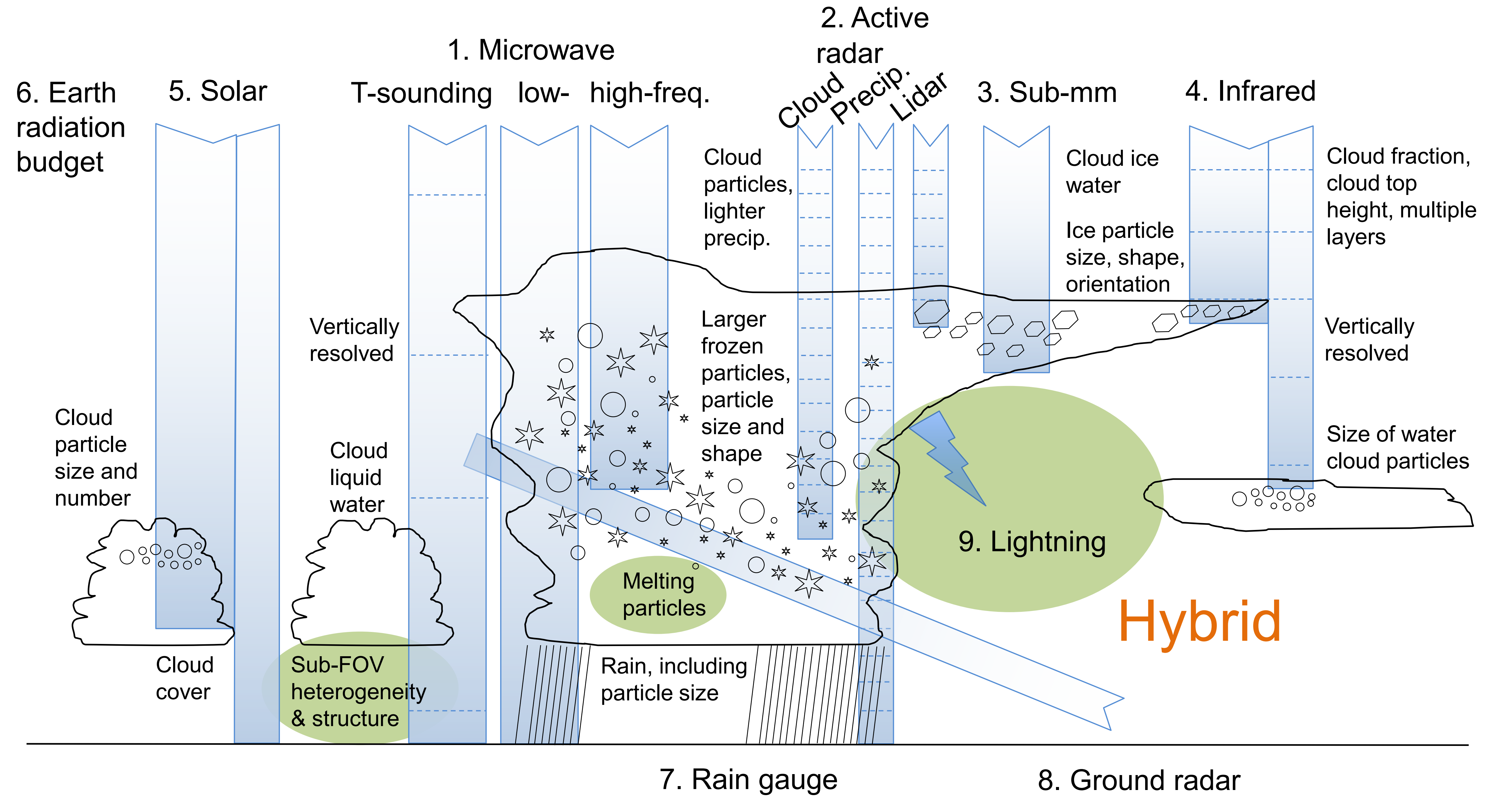
What actually is "ground truth"?

- Radar reflectivity is **not** rain
  - Reflectivity is affected by shape & melting layer effects, particle size distribution, particle orientation, attenuation along the beam (other rain events)
- Rain at radar altitude is **not** rain at the surface
  - Evaporation
  - Small scale wind effects / turbulence (up and down-drafts, gusts)
- Rain gauges provide **observations**, not truth
  - "So, how much of the earth's surface area **is** covered by rain gauges?", Kidd et al. (2017, https://doi.org/10.1175/BAMS-D-14-00283.1):
    - Earth's surface area: **~510,000,000,000,000 m^2**
    - GTS rain gauge orifice surface area: **295 m^2** (just bigger than a tennis court)
  - … and don't forget wind-driven "undercatch" errors.
- **There is no ground truth**

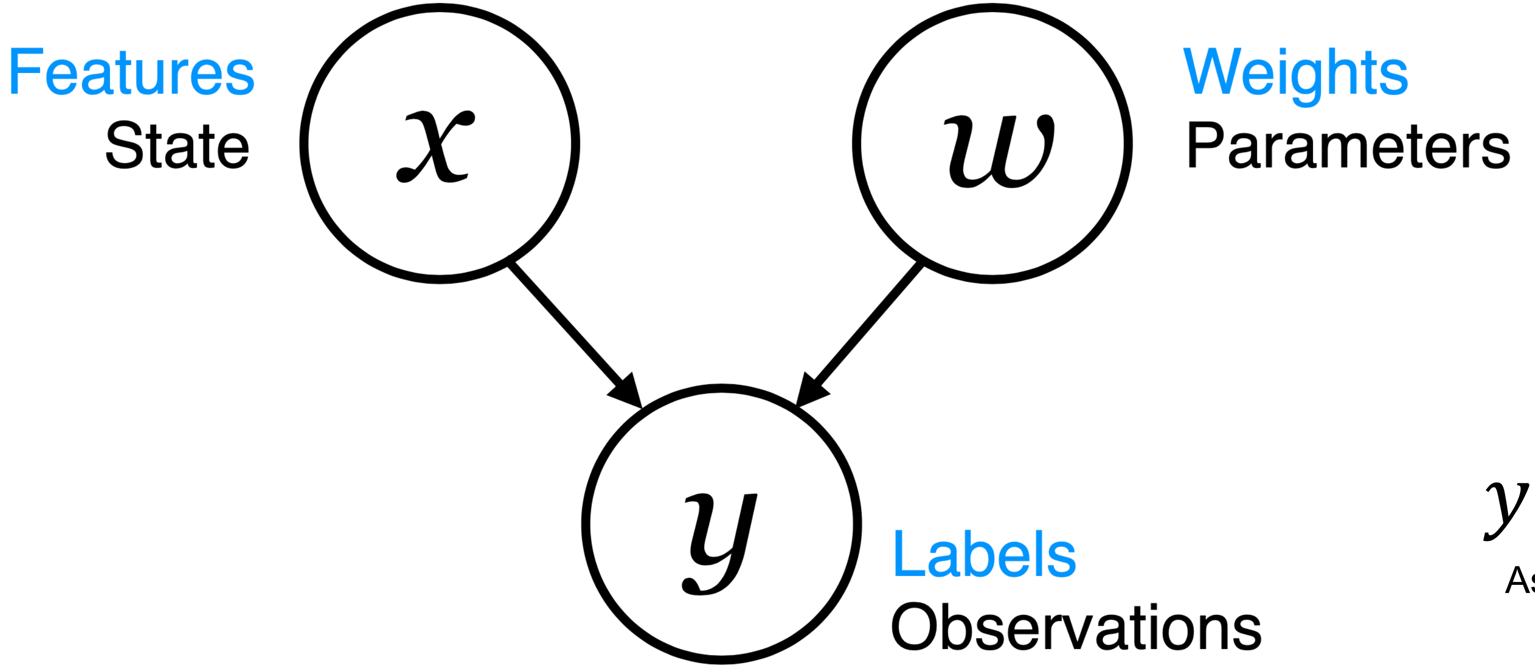


6. Earth radiation budget
5. Solar
1. Microwave
T-sounding
low-
high-freq.
2. Active radar
Cloud
Precip.
Lidar
3. Sub-mm
4. Infrared
Cloud particles, lighter precip.
Cloud ice water
Ice particle size, shape, orientation
Cloud fraction, cloud top height, multiple layers
Vertically resolved
Larger frozen particles, particle size and shape
Vertically resolved
Size of water cloud particles
Cloud particle size and number
Cloud liquid water
9. Lightning
Melting particles
Cloud cover
Sub-FOV heterogeneity & structure
Rain, including particle size
7. Rain gauge
8. Ground radar


6. Earth radiation budget
5. Solar
1. Microwave
T-sounding
low-
high-freq.
2. Active radar
Cloud
Precip.
Lidar
3. Sub-mm
4. Infrared
Latent space
9. Lightning
7. Rain gauge
8. Ground radar



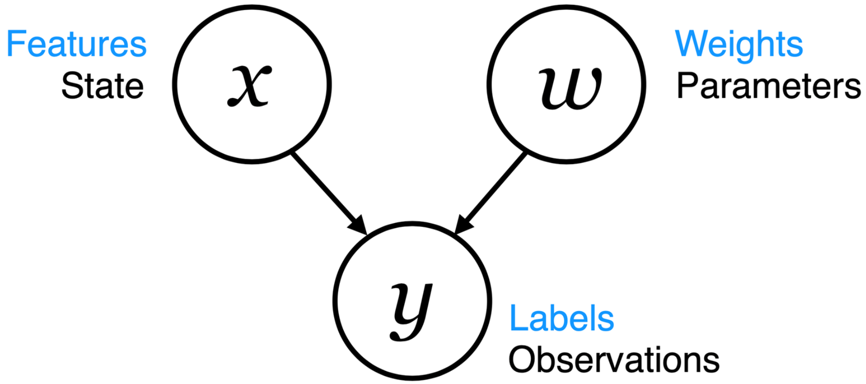
2. Active radar
1. Microwave
6. Earth radiation budget
5. Solar
T-sounding
low-
high-freq.
Cloud
Precip.
Lidar
3. Sub-mm
4. Infrared
Cloud particles, lighter precip.
Cloud ice water
Ice particle size, shape, orientation
Cloud fraction, cloud top height, multiple layers
Vertically resolved
Larger frozen particles, particle size and shape
Vertically resolved
Cloud particle size and number
Cloud liquid water
Size of water cloud particles
9. Lightning
Melting particles
Hybrid
Cloud cover
Sub-FOV heterogeneity & structure
Rain, including particle size
7. Rain gauge
8. Ground radar



$$y = h(x, w)$$

As a Bayesian network

Bayes' theorem



$$P(y, x, w)$$

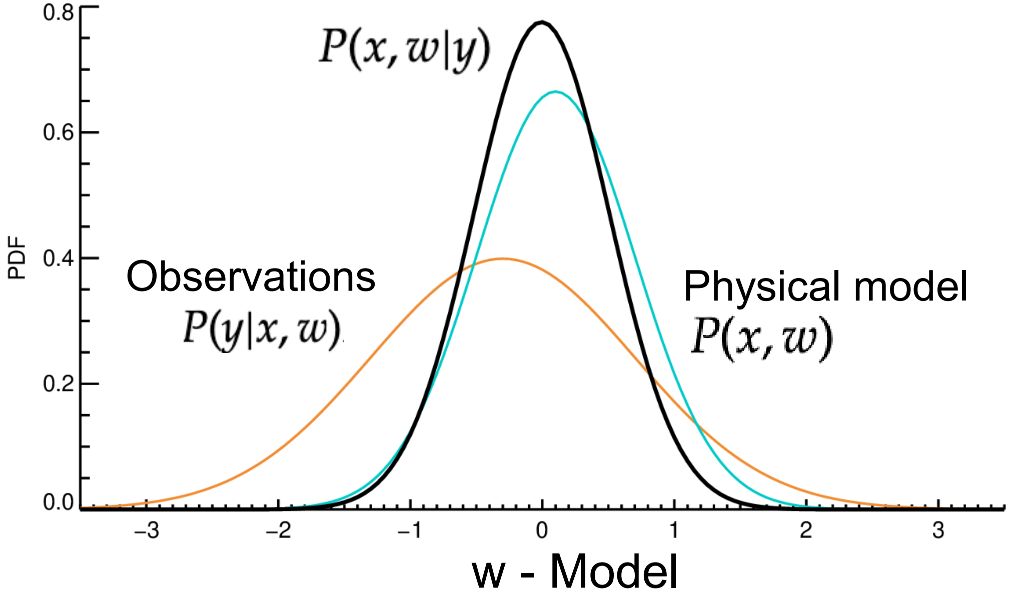
Posterior knowledge of state and model	Observations (likelihood function)	Prior knowledge of state and model

$$P(x, w|y) = \frac{P(y|x, w)P(x, w)}{P(y)}$$

Machine learning vs. DA

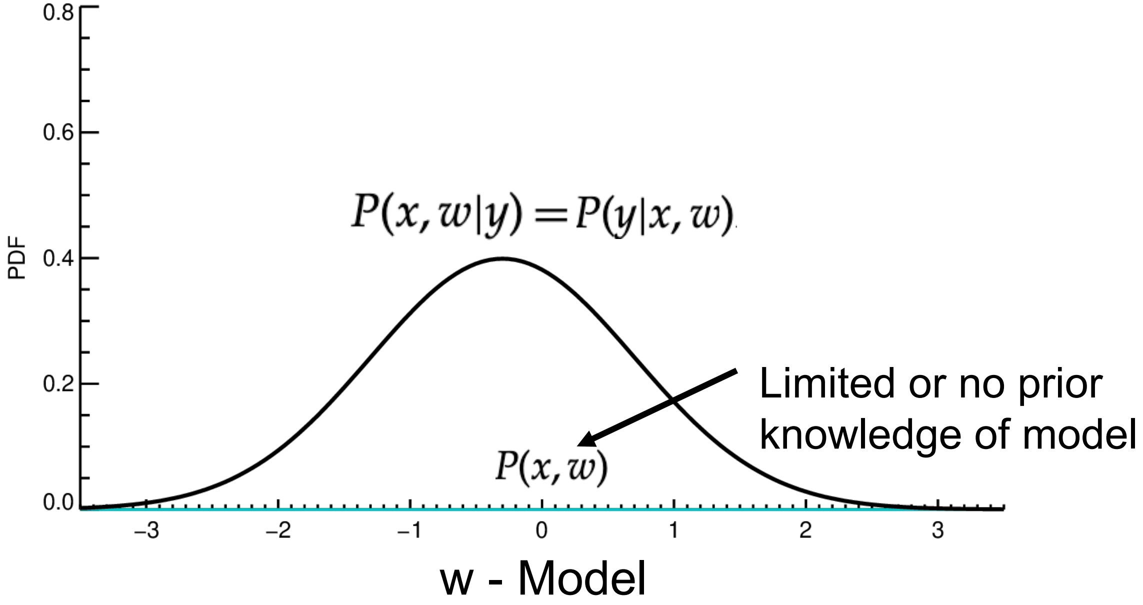
$$P(x,w|y) = \frac{P(y|x,w)P(x,w)}{P(y)}$$

Data assimilation

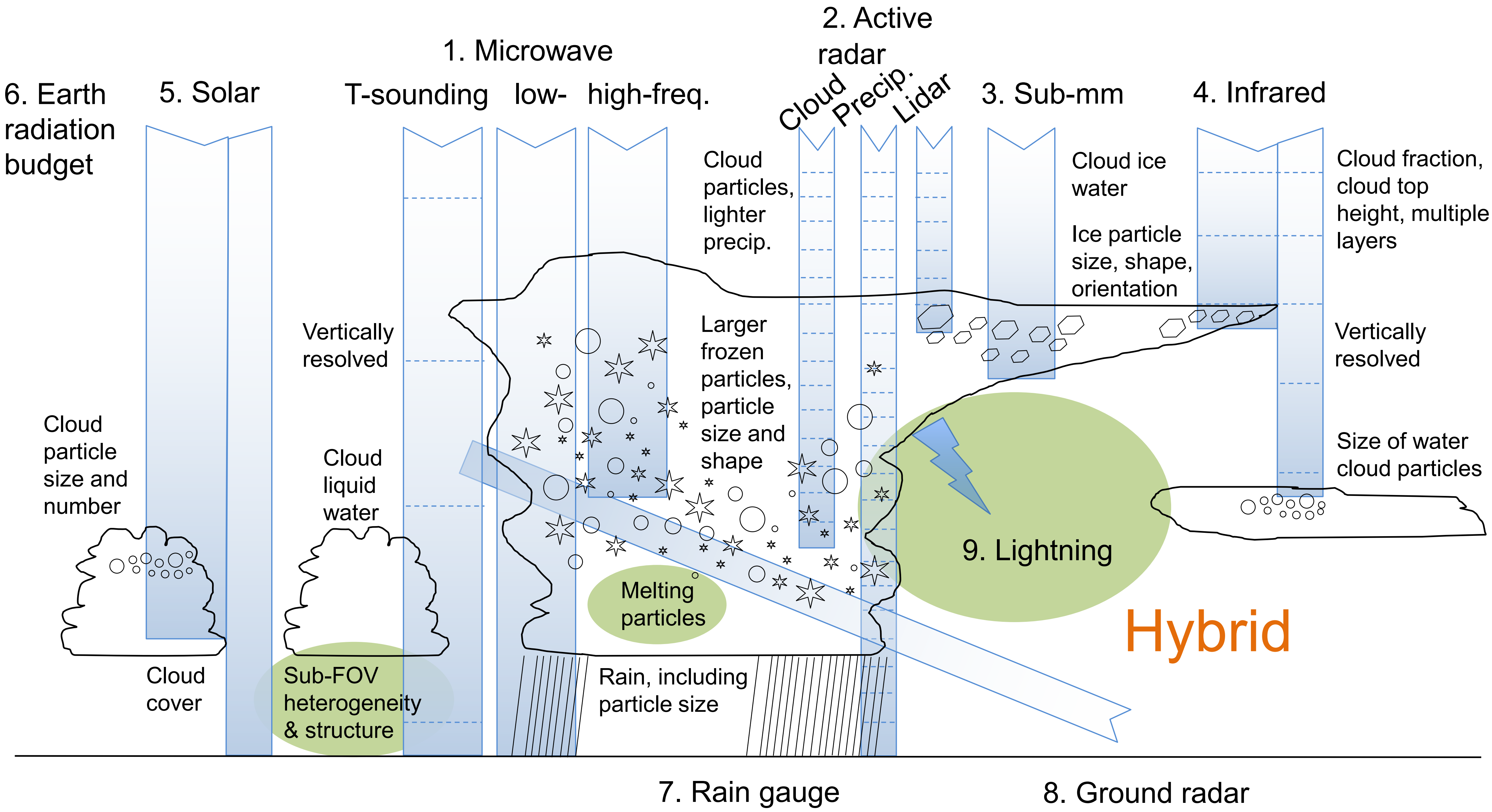


Our knowledge of the model is better when we combine physics and observations, even if the physical knowledge is poor

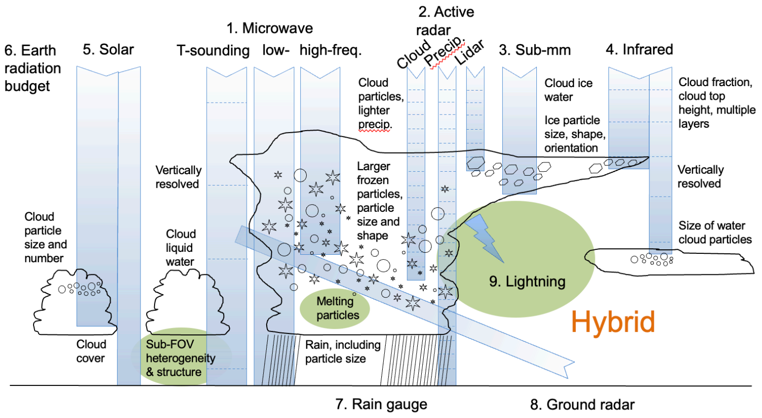
Machine learning



Our knowledge of the model is no better than what we know from observations



1. Microwave
T-sounding
low-
high-freq.
2. Active radar
Cloud
Precip.
Lidar
3. Sub-mm
4. Infrared
5. Solar
6. Earth radiation budget
7. Rain gauge
8. Ground radar
9. Lightning
Hybrid
Cloud particles, lighter precip.
Cloud ice water
Ice particle size, shape, orientation
Cloud fraction, cloud top height, multiple layers
Vertically resolved
Size of water cloud particles
Larger frozen particles, particle size and shape
Vertically resolved
Cloud particle size and number
Cloud liquid water
Melting particles
Cloud cover
Sub-FOV heterogeneity & structure
Rain, including particle size



Melting particles

Mass, shape, PSD, fall speed and subgrid heterogeneity of frozen particles

Temperature, humidity, winds

Mass, shape, PSD, fall speed and subgrid heterogeneity of melting particles

Model choice eg :
- Physical constraints (equations)
- Empirical processes (ML)

Prior constraints (background errors)

Bulk optical properties of melting particles (e.g. radar reflectivity)

Model choice eg: a fully empirical trainable mapping (ML)

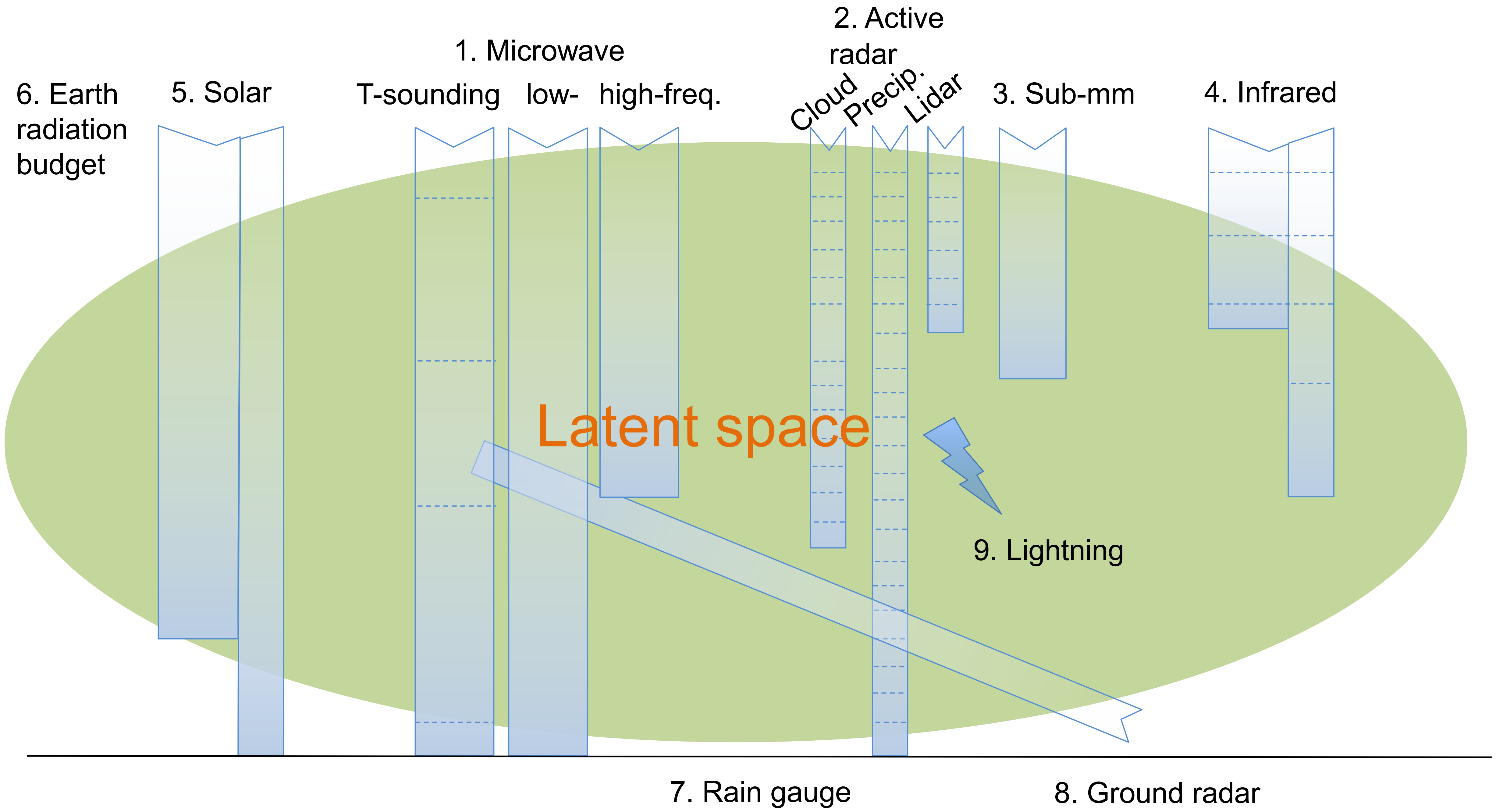
Prior constraints (background errors)

How to build this “granular hybrid” approach

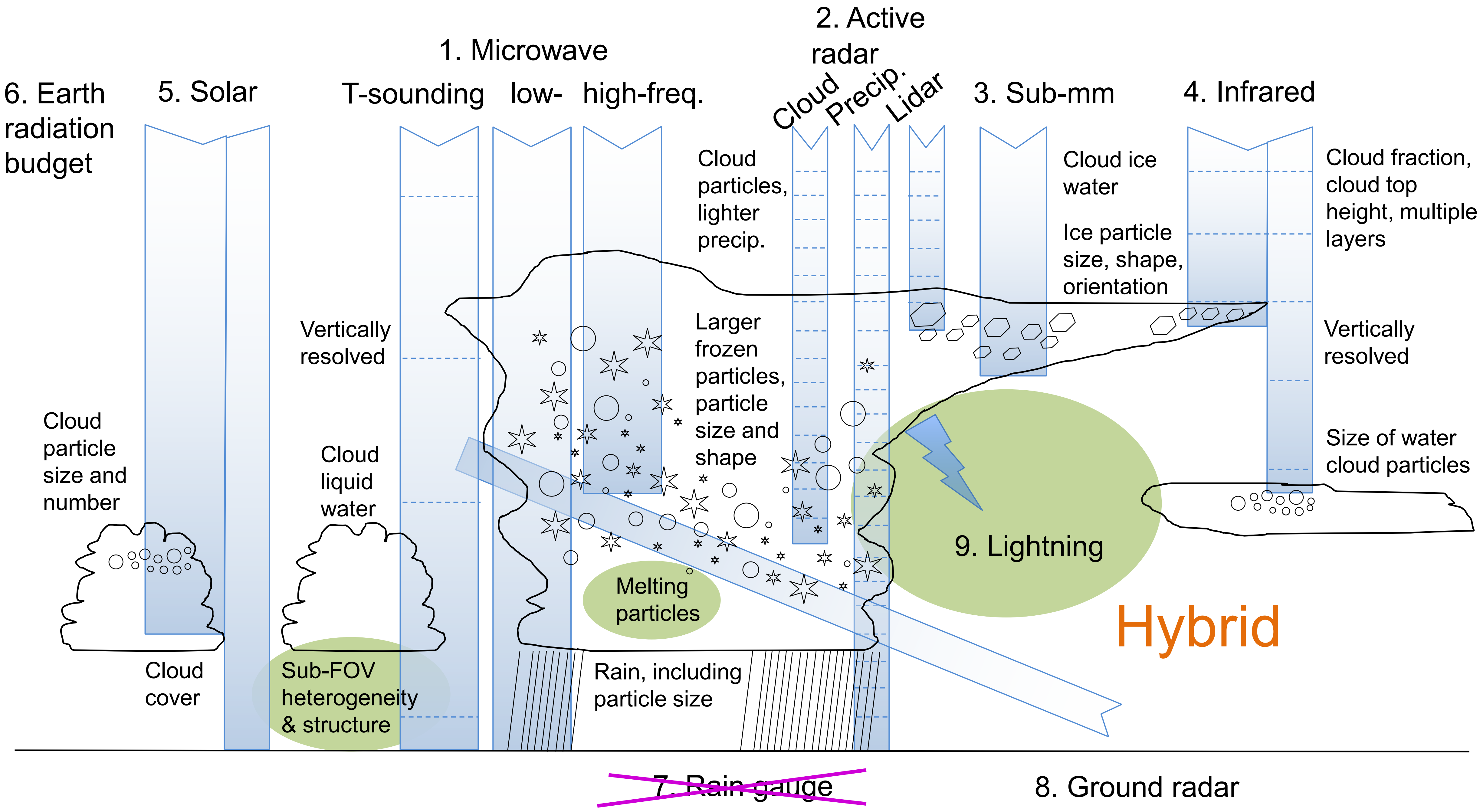
- For each conditionally independent PDF (model fragment) we need to describe:
  - Current best state estimate
    - And our knowledge / confidence in this (PDF)
  - Current model
    - Physical equations, and our confidence in them
    - Empirical components (ML) and our confidence in them
  - Gradients with respect to all variables
    - And ideally also with respect to the prior PDFs, which will themselves be uncertain (see e.g. Dee and Da Silva, 1999, https://doi.org/10.1175/1520-0493(1999)127<1822:MLEOFA>2.0.CO;2)
- We need a model (or graph) to join all these fragments together, compute gradients, make forecasts
- We apply data assimilation based on all available observations

Granular hybrid ML-DA methods: how?

- Add empirical components into existing DA systems (including online fine tuning, e.g. Marcin & Alban’s talk).
  - This is slow and difficult work (e.g. sea ice, to follow)
  - Re-training is going to be very important (e.g. out of dataset issues)
- This is currently much easier to implement in ML frameworks than in DA
- **Long term, granular hybrids need a whole new technical infrastructure**
  - **A new hybrid of pyTorch / Tensorflow / OOPS / IFS etc.**



2. Active radar
1. Microwave
6. Earth radiation budget
5. Solar
T-sounding
low-
high-freq.
Cloud
Precip.
Lidar
3. Sub-mm
4. Infrared
Latent space
9. Lightning
7. Rain gauge
8. Ground radar



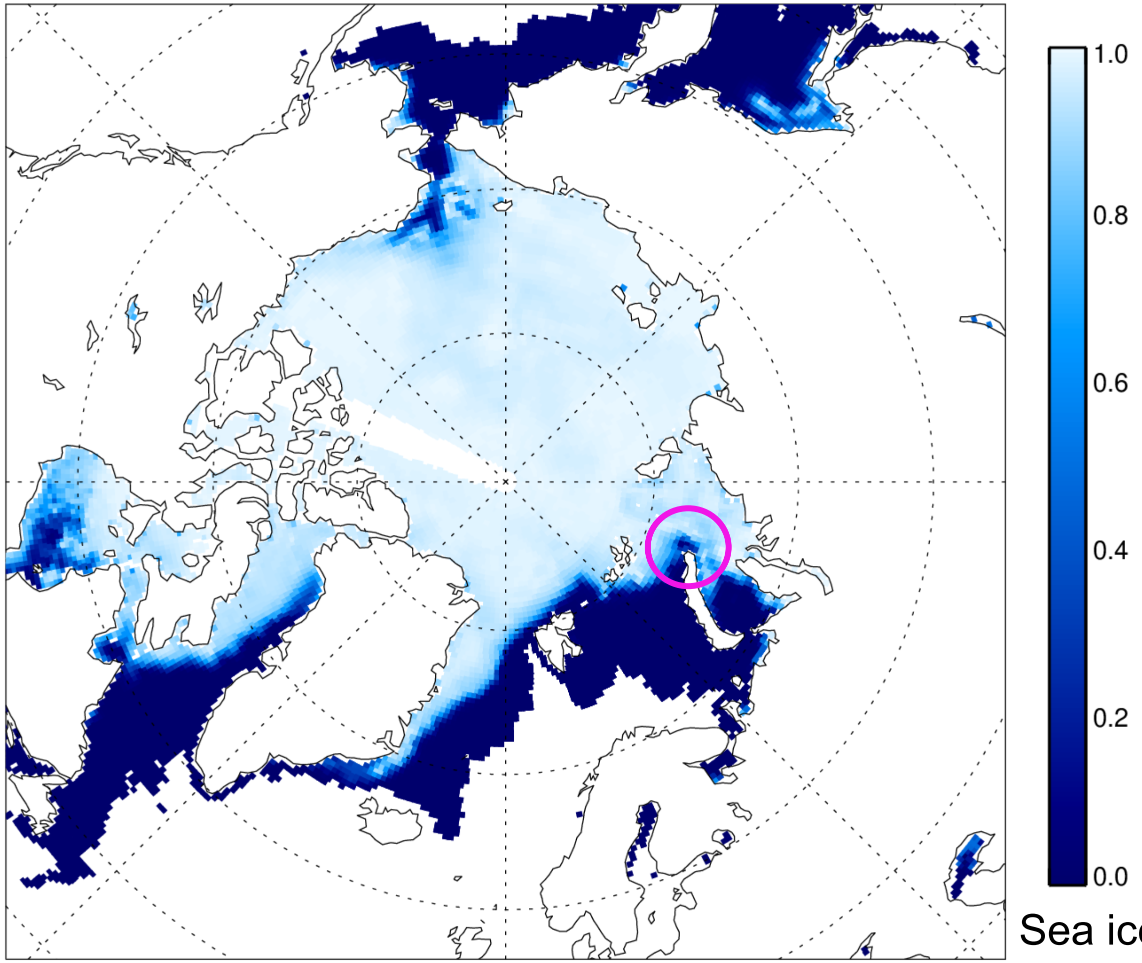
6. Earth radiation budget
5. Solar
1. Microwave
T-sounding
low-
high-freq.
2. Active radar
Cloud
Precip.
Lidar
3. Sub-mm
4. Infrared
Cloud particles, lighter precip.
Cloud ice water
Ice particle size, shape, orientation
Cloud fraction, cloud top height, multiple layers
Vertically resolved
Larger frozen particles, particle size and shape
Vertically resolved
Cloud particle size and number
Cloud liquid water
Size of water cloud particles
9. Lightning
Melting particles
Hybrid
Cloud cover
Sub-FOV heterogeneity & structure
Rain, including particle size
~~7. Rain gauge~~
8. Ground radar

What if there are also no direct observations?

- GraphDOP Arxiv paper (Alexe et al., 2024):
  - “One obvious limitation of using observations as targets in the training is that predictions can only be made for physical variables for which we have direct observations. For example, although the results in this paper showed accurate predictions of sea ice in radiance space, it is not possible to make predictions for more abstract quantities, such as sea ice concentration, without having direct observations of that variable.”

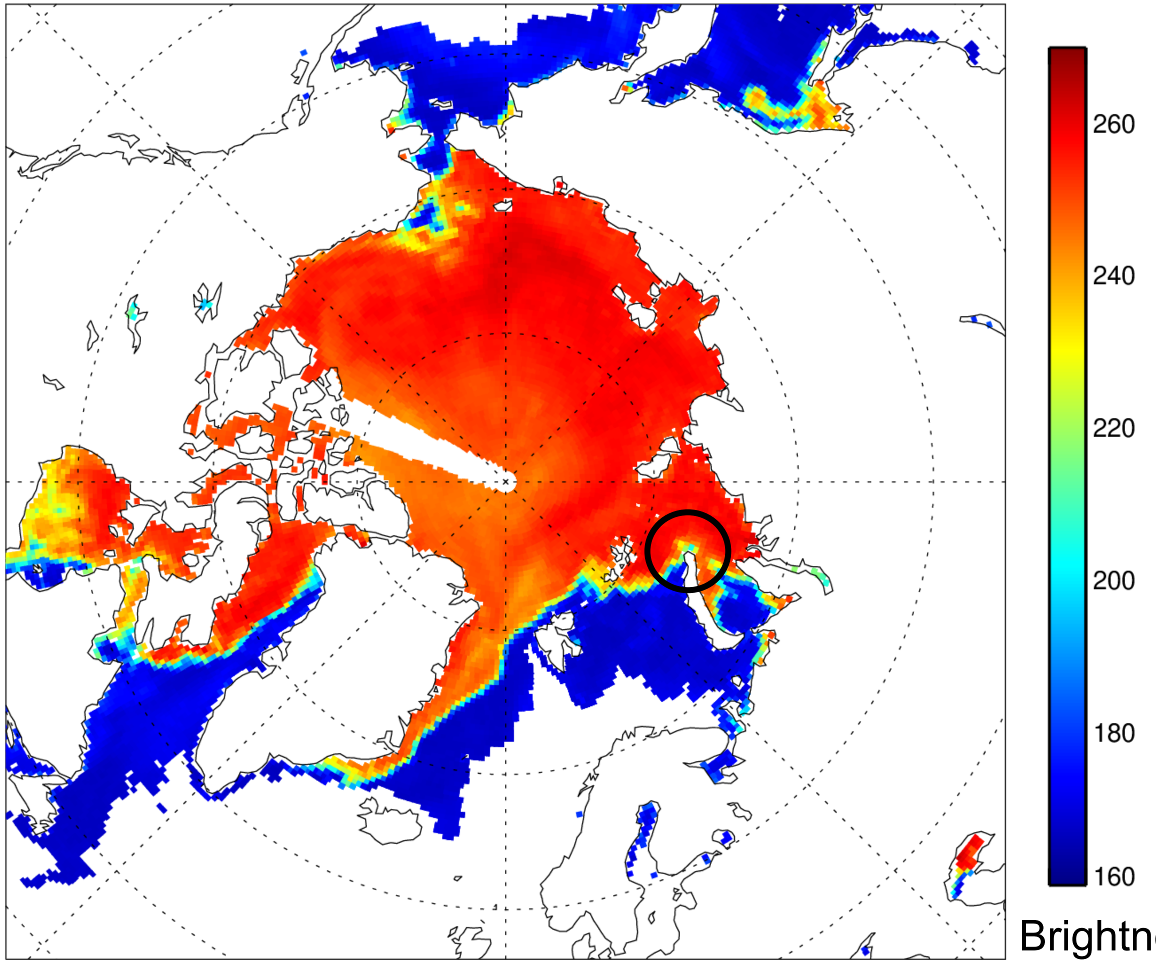
How to improve this, given this?

NEMO/SI3 background, 00Z 10th Dec 2022

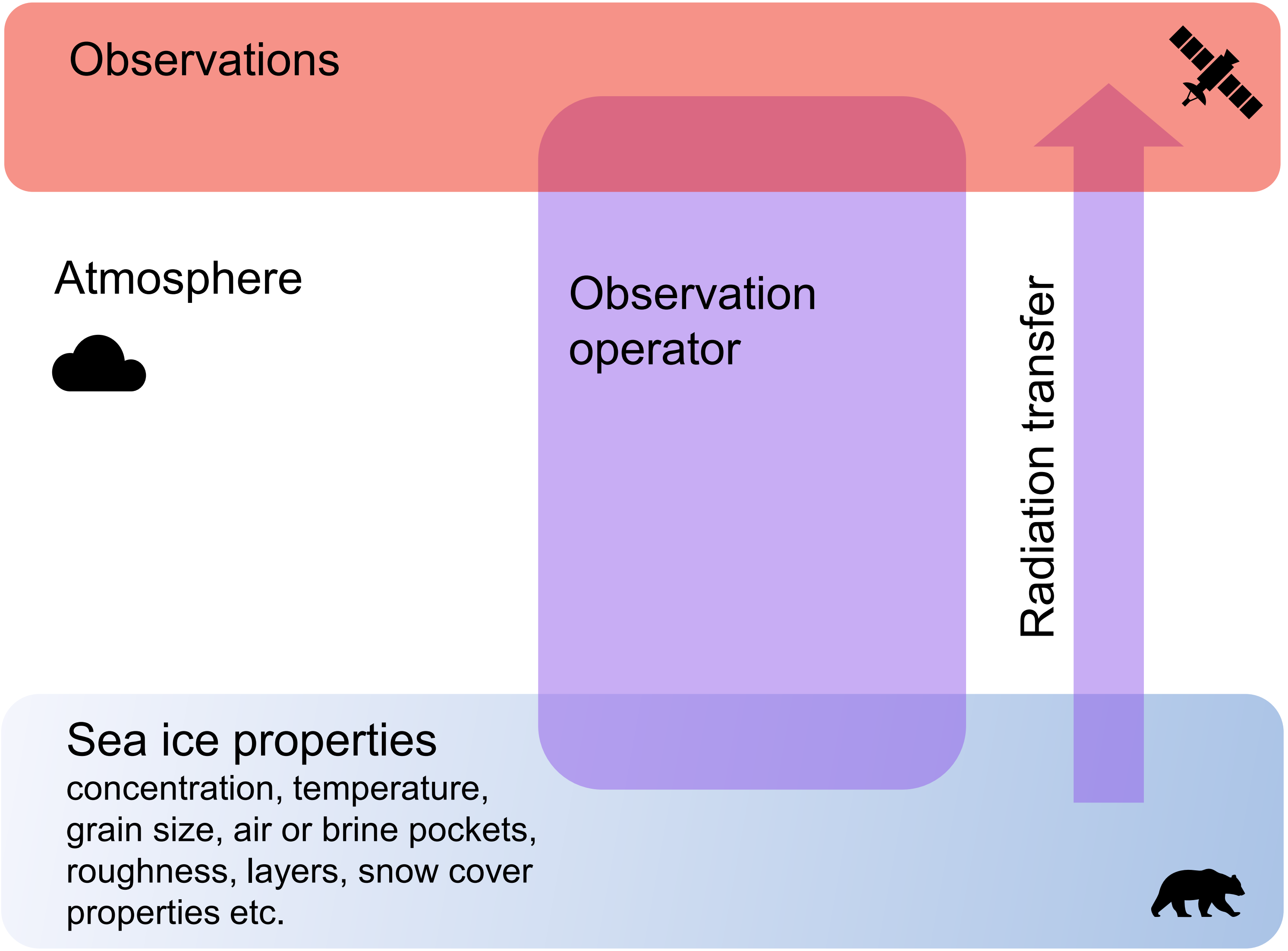


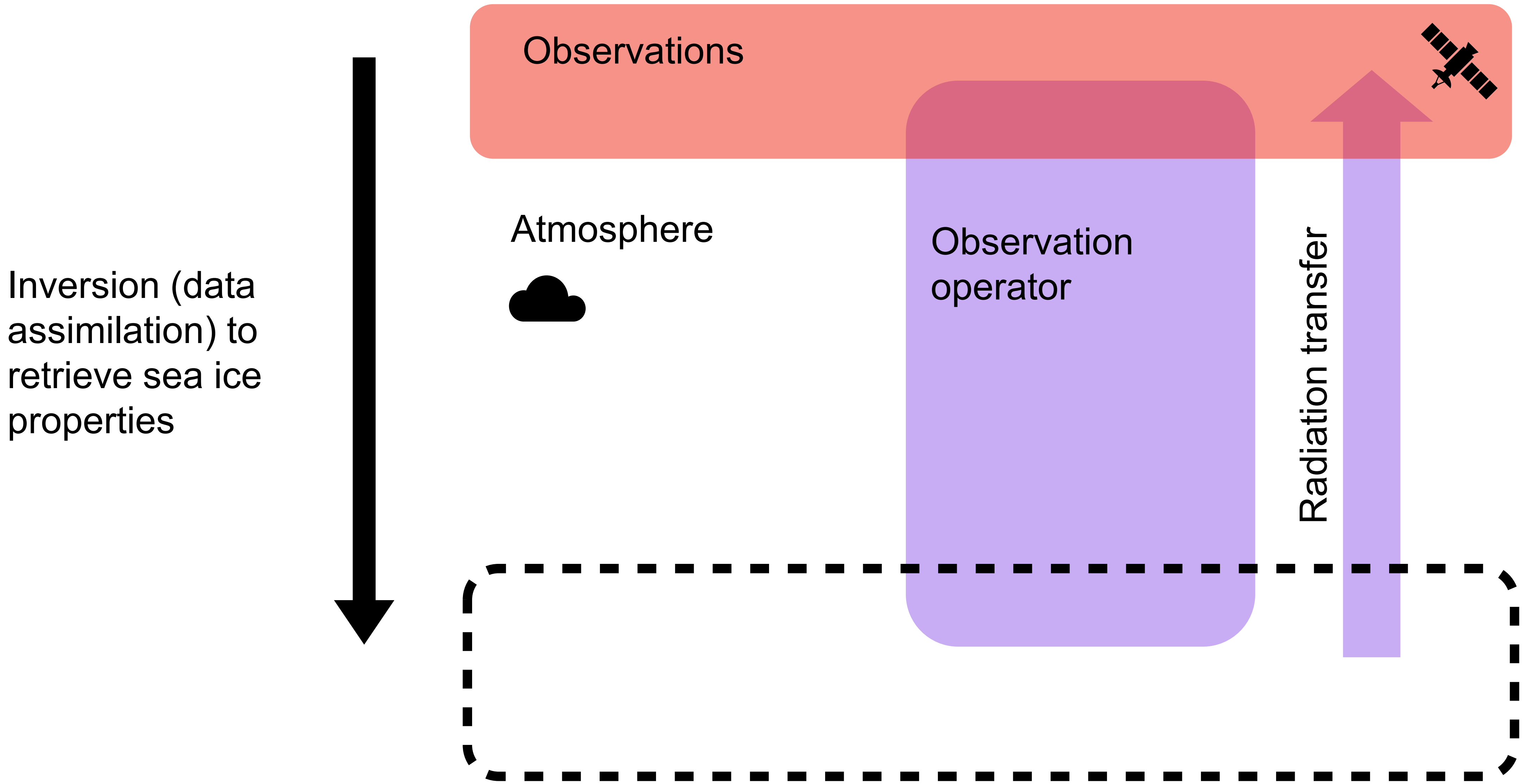
Sea ice concentration

AMSR2 10 GHz observation, 00Z 10th Dec 2022

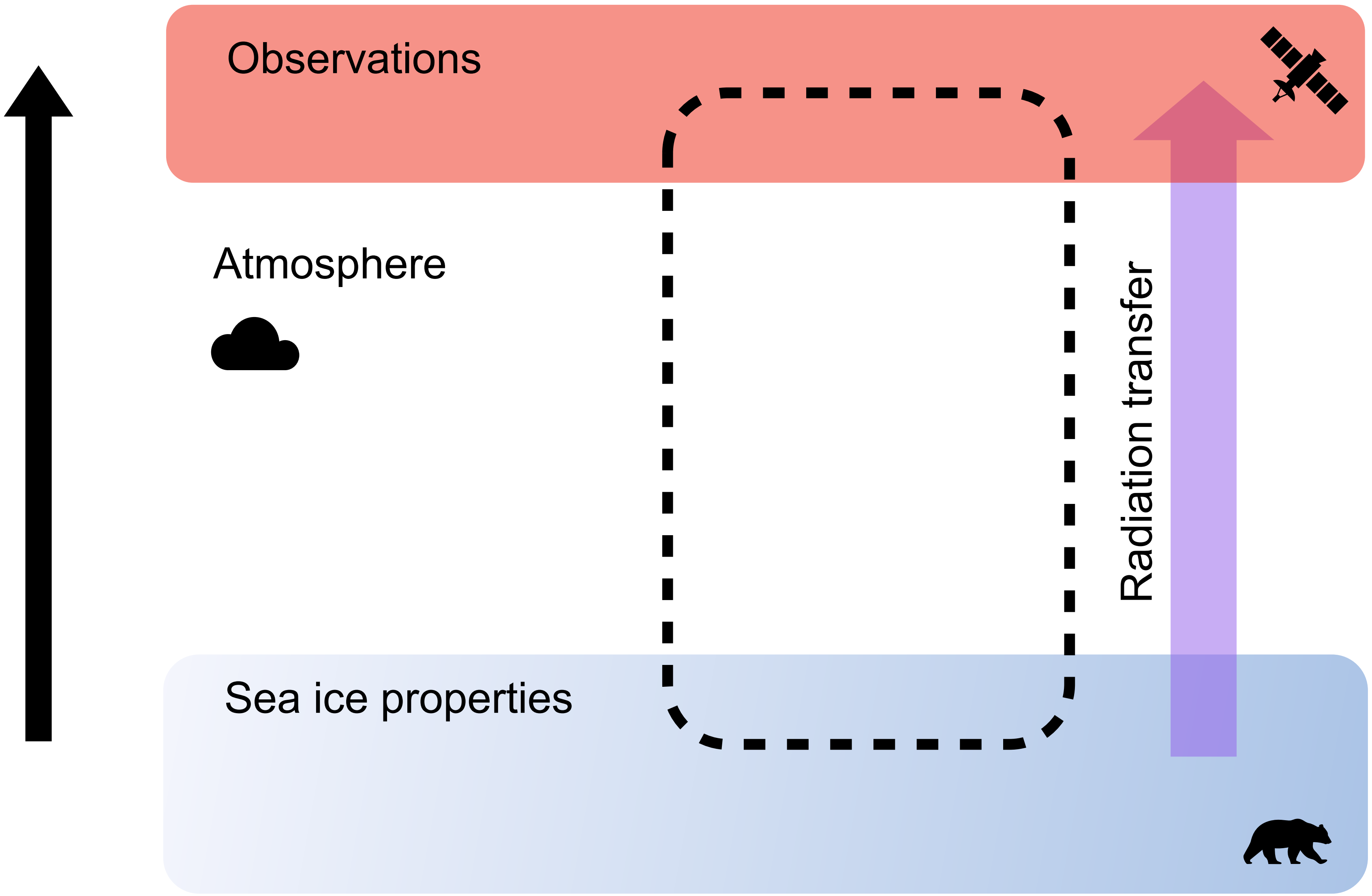


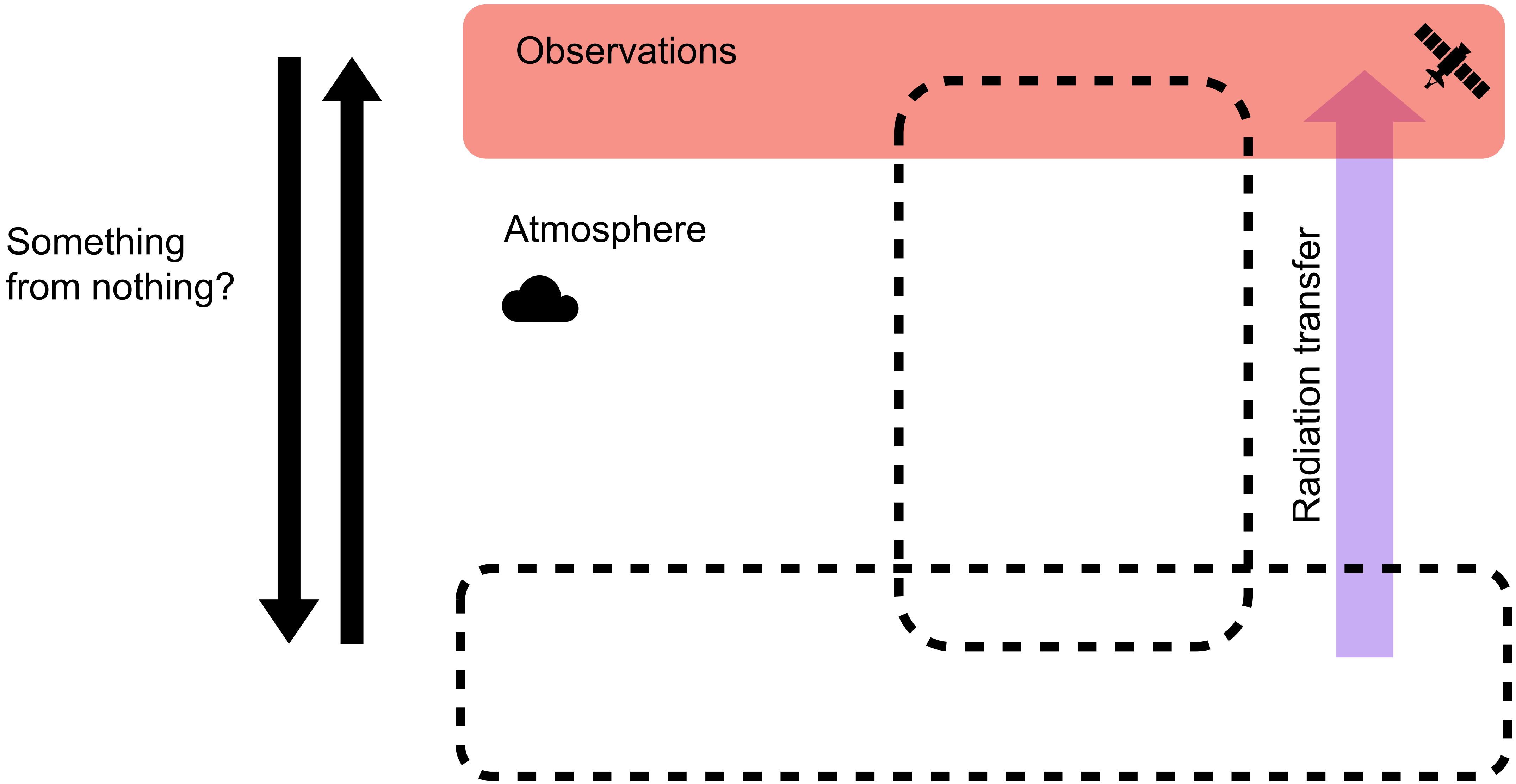
Brightness temperature


Observations
Atmosphere
Observation operator
Radiation transfer
Sea ice properties
concentration, temperature, grain size, air or brine pockets, roughness, layers, snow cover properties etc.

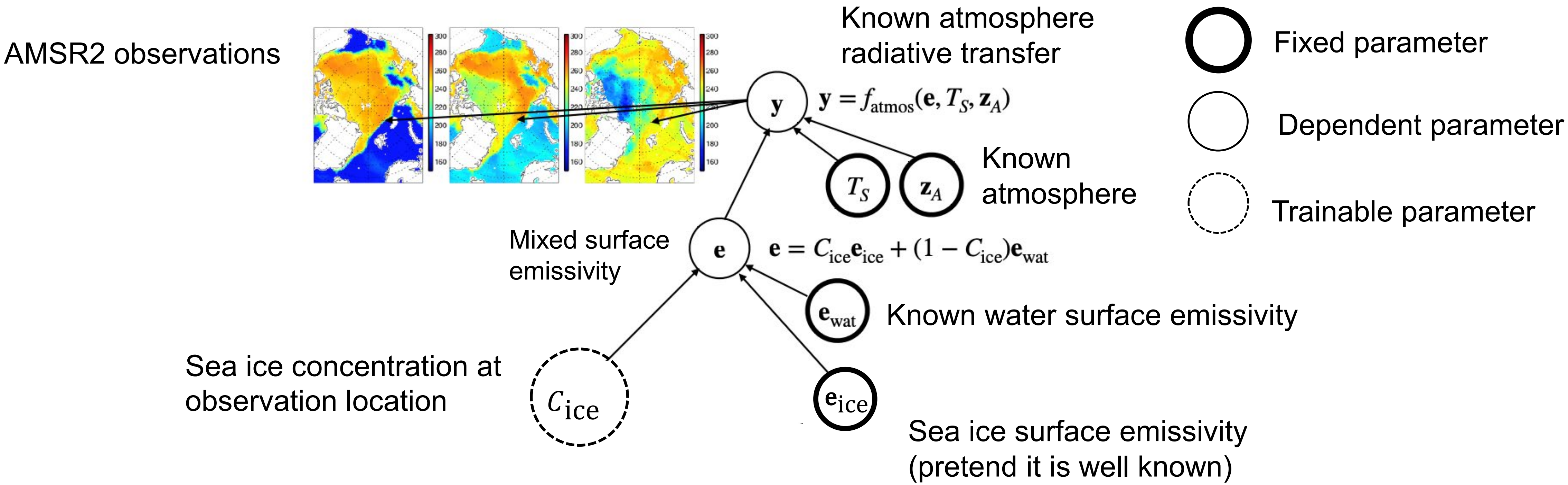

Observations
Atmosphere
Observation operator
Radiation transfer
Inversion (data assimilation) to retrieve sea ice properties

Machine learning to find the observation operator

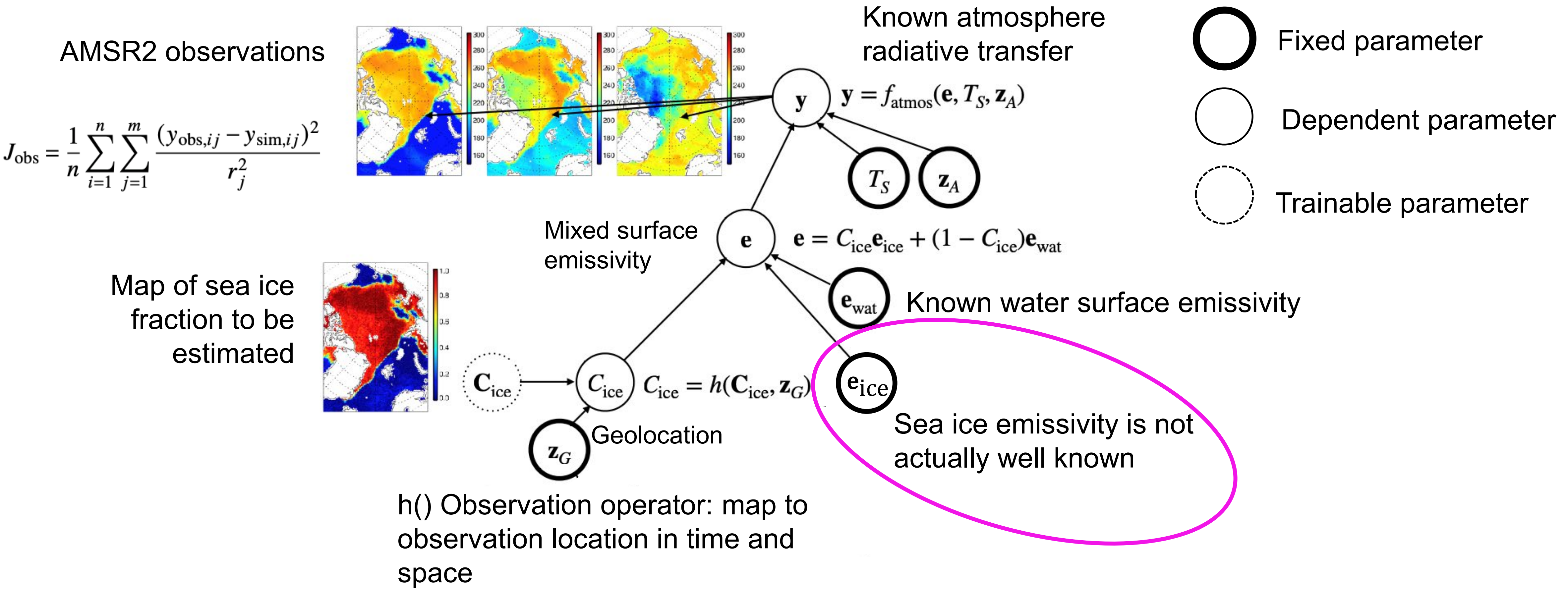



Something
from nothing?
Observations
Atmosphere
Radiation transfer

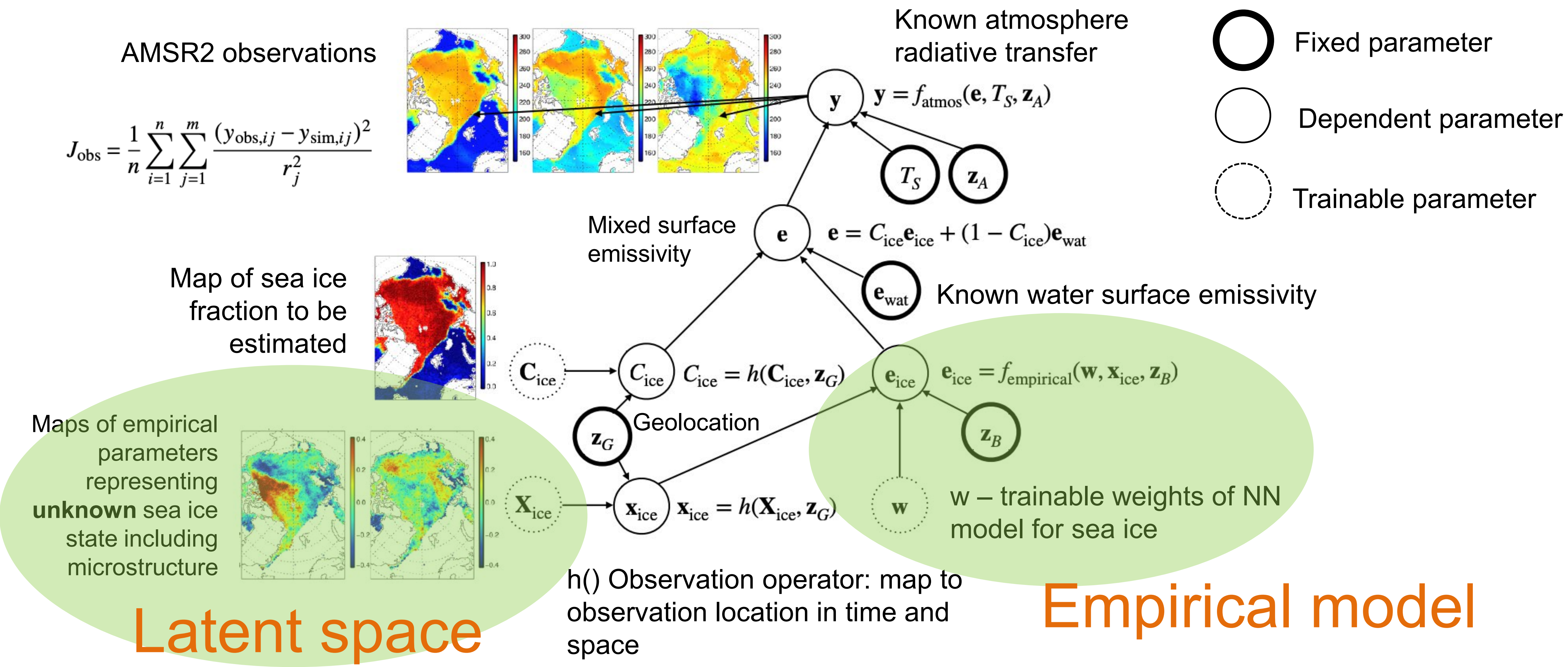
Physical (Bayesian) network representation of sea ice and snow radiative transfer for variational data assimilation



Physical (Bayesian) network representation of sea ice and snow radiative transfer for variational data assimilation



The whole trainable empirical-physical network

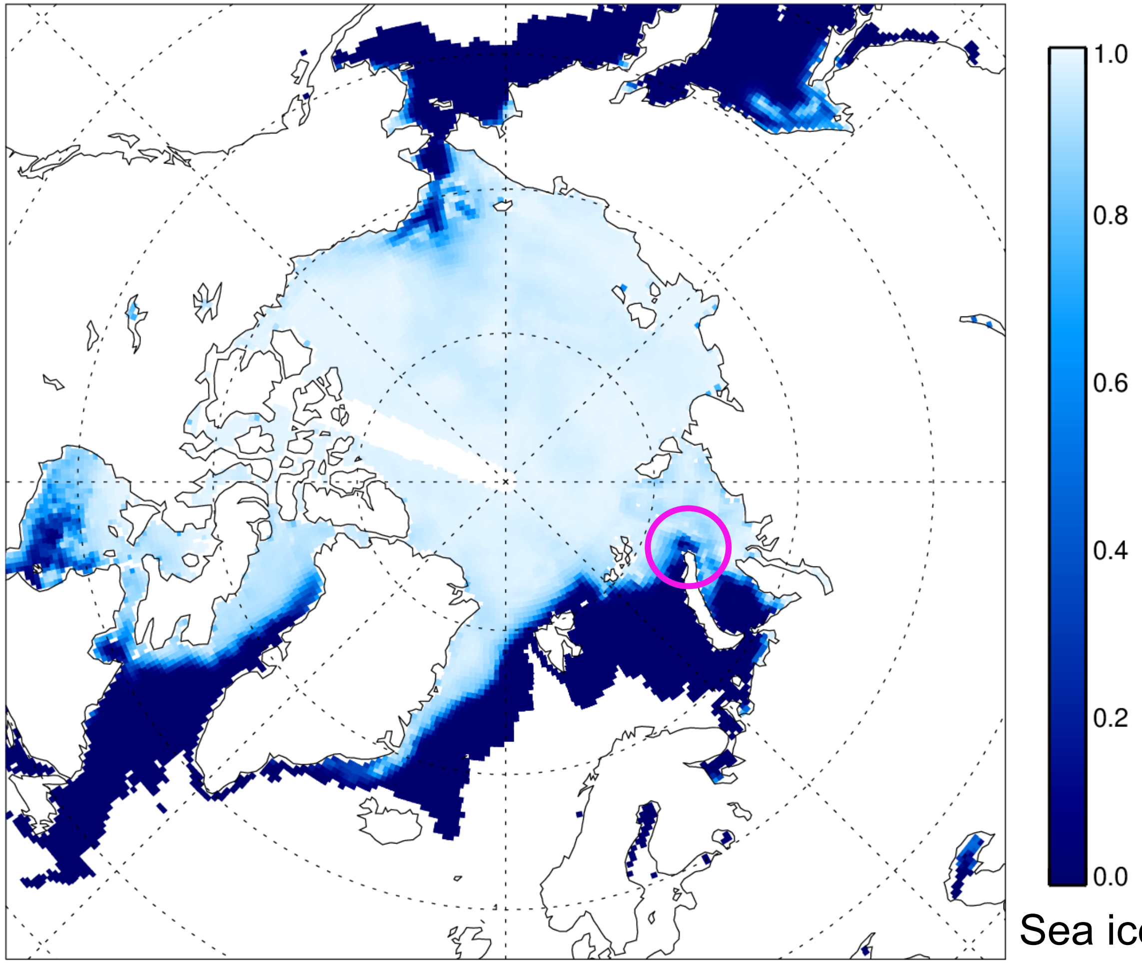


Given this offline-trained hybrid network, how to include sea ice assimilation in weather forecasting?

- Train the hybrid network against a year (or more) of observations outside the main DA framework
- Implement the trained network into 4D-Var data assimilation for the atmosphere (currently operational at cycle 49r1, since Nov 2024)
  - Activate assimilation of microwave window channels over sea ice for the first time
  - At each observation location, retrieve SIC and empirical sea ice state variables
- Pass the retrieved SIC as a pseudo-observation to the ocean data assimilation component (NEMOVAR) via outer-loop coupling (intended to be operational in cycle 50r1, Nov 2025)
- Further reading:
  - https://doi.org/10.1002/qj.4797
  - https://doi.org/10.1029/2023MS004080

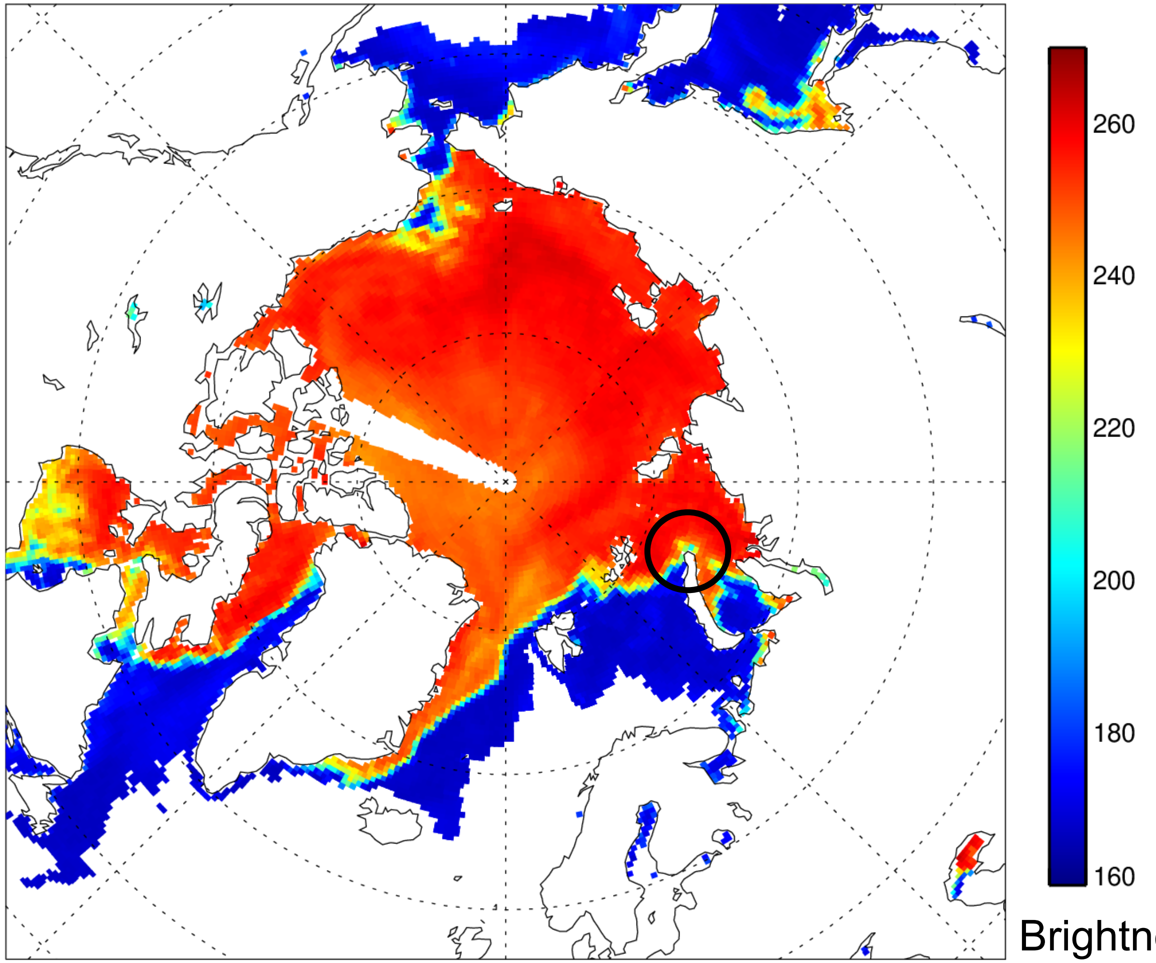
How to improve this, given this?

NEMO/SI3 background, 00Z 10th Dec 2022



Sea ice concentration

AMSR2 10 GHz observation, 00Z 10th Dec 2022

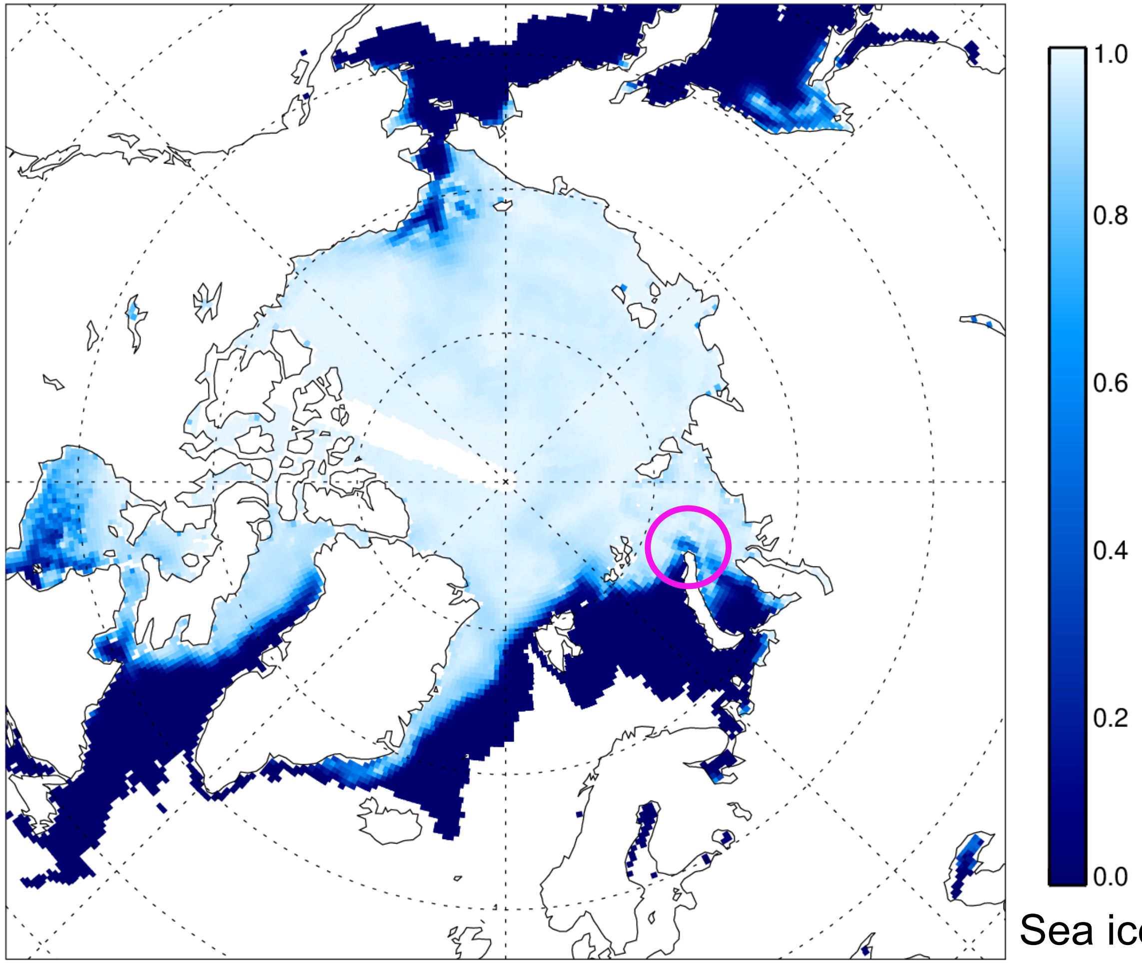


Brightness temperature

Analysis and increments in SIC from NEMOVAR

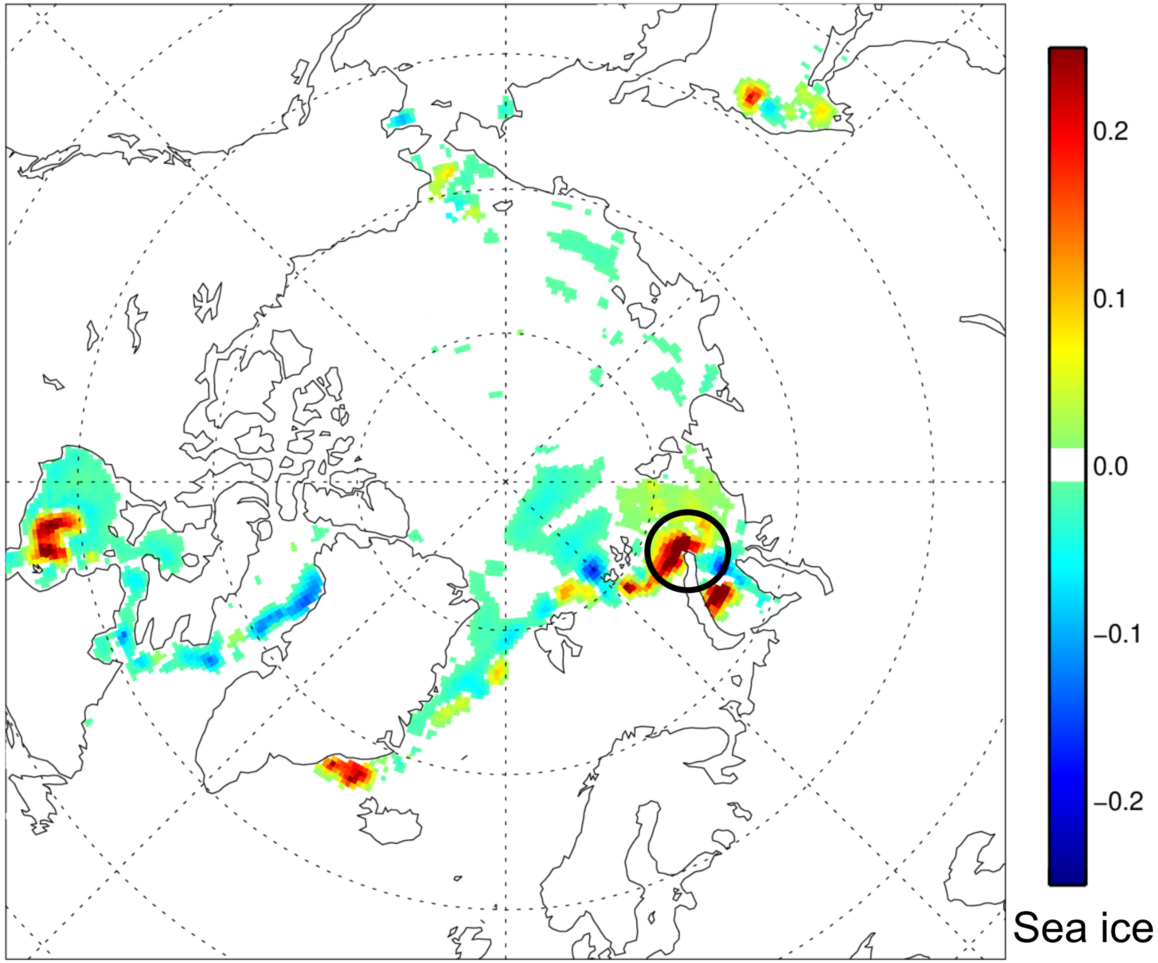
Thanks Phil Browne!

NEMO/SI3 analysis, 00Z 10th Dec 2022



Sea ice concentration

SIC increment, 00Z 10th Dec 2022



Sea ice concentration

Reasons to improve physical DA and forecasting using ML

- ML suggests gains of at least 1 forecast day are possible
- Traditional model development processes and parameter estimation have not fully succeeded in improving physical models
- Some aspects of the physics are not well-known (e.g. melting snow particles, sea ice microstructure)

Reasons to improve ML-based DA and forecasting using physical DA

- Standard ML operates within the space of analysed datasets (e.g. ERA5)
- ML does not provide the intermediate variables (e.g. size of hail), just latent variables
- DOP-type ML (pure observations) could break out of analysis-space, but cannot provide target variables that are not directly observed (e.g. sea ice)

Hybrid ML-DA methods

There is no ground truth, only data assimilation

Bayes' thereom shows that a combination of prior knowledge (e.g. physics) with observations provides more knowledge than either alone