

Error Decomposition of SWH Estimates

Haoyu Jiang (Haoyujiang@szu.edu.cn)

Department of
Oceanography, Shenzhen
University, Shenzhen, China

IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 61, 2023

4205613

Random, Environmental, and Representativeness
Errors in Ocean Remote Sensing Versus
In Situ Data: An Example of Wave
Heights From Altimeters

Haoyu Jiang 

Error Decomposition: Definition of Error Components

Given a SWH in a given spatial-temporal range, the measurements (X) of the truths (T) from a given data source can be decomposed as

$$X = T + \sigma + e + r$$

σ : random error

e : environmental error

r : representativeness error

First, the spatial-temporal range of T needs to be clearly defined.

We can arbitrarily define one, but we need to know what is the exact thing we want to estimate.

Temporal “average” at a given location [Buoy]

Spatial “average” at a give time point [Satellite]

From space-time “average” to a space-time point, can be explained in many ways [Model]

There difference is a part of the representativeness error (discuss later)

Error Decomposition: Definition of Error Components

Given a SWH in a given spatial-temporal range, the measurements (X) of the truths (T) from a given data source can be decomposed as

$$X = T + \sigma + e + r$$

**σ : random error : expected value is 0 , uncorrelated to any variables
“purely random” part of the data**

$$\langle \sigma_a \rangle = \langle \sigma_a T \rangle = \langle \sigma_a e_a \rangle = \langle \sigma_a r_a \rangle = \langle \sigma_a \sigma_b \rangle = \langle \sigma_a e_b \rangle = \langle \sigma_a r_b \rangle = 0$$

subscripts a and b can be any two data sources: e.g., Altimeter and Buoy

When If N independent estimates are available for the same truth, the variance can be reduced from $\langle \sigma^2 \rangle$ to $\langle \sigma^2 \rangle / N$. **【 Central Limit Theorem 】**

Error Decomposition: Definition of Error Components

Given a SWH in a given spatial-temporal range, the measurements (X) of the truths (T) from a given data source can be decomposed as

$$X = T + \sigma + e + r$$

e : environmental error

Error components that are dependent on other variables

Can be expressed as $e = f(V)$ where V can be any variable

Example: When a system is not well-calibrated, we can use $e = kX + b$ to reduce this error (linear calibration)

Sometimes there are unknown processes/variables impact this error term.

[A bad example: Maybe wave breaking can lead to slightly overestimates of SWH?

I don't know....]



Error Decomposition: Definition of Error Components

Given a SWH in a given spatial-temporal range, the measurements (X) of the truths (T) from a given data source can be decomposed as

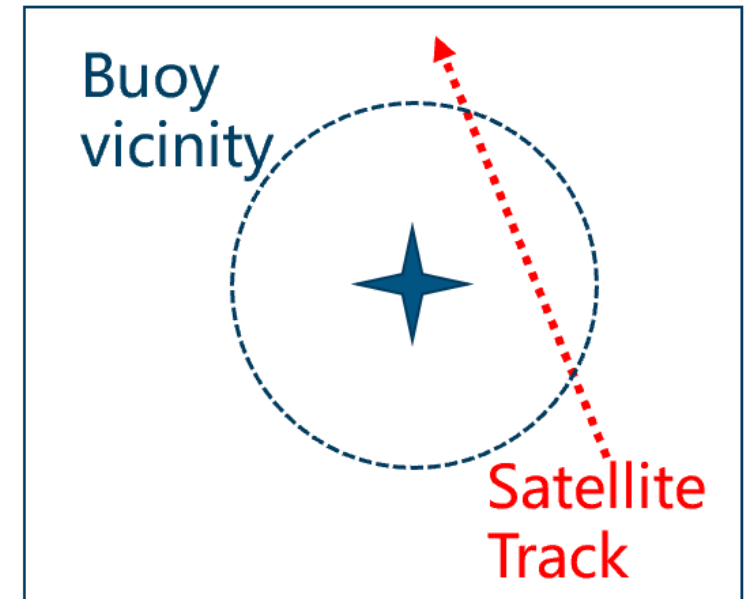
$$X = T + \sigma + e + r$$

r : representativeness error

Errors due to the fact that different sensors/models or different measurements from the same sensor are estimating different truth.

Even for two remote sensing datasets, it is impossible to obtain two measurements covering exactly the same area at the exactly same time.

So again, when the word "truth" is used, it is necessary to give an explicit definition.



Error Decomposition: Definition of Error Components

Given a SWH in a given spatial-temporal range, the measurements (X) of the truths (T) from a given data source can be decomposed as

$$X = T + \sigma + e + r$$

σ : random error

e : environmental error

r : representativeness error

The three error components are not strict with respect to the definition.

e.g., the bias between buoy and altimeter SWH when buoys are too close to land can be regarded as either environmental error (systematic) or representativeness error (due to different locations)

But they might still be helpful for the understanding, analysis, and suppression of errors.

Double & Triple Collocation

Double Collocation

$$B = T + \sigma_B + e_B + r_B = T + \sigma_B \quad B : \text{Buoy}$$

$$R = T + \sigma_R + e_R + r_R \quad R : \text{RS}$$

Truth: 20-min sampled SWH at buoy's locations

Buoy: Reference data, without environment error and representativeness error

The (MSE) error between buoy and RS is then:

$$\langle (B - R)^2 \rangle = \langle \sigma_B^2 \rangle + \langle \sigma_R^2 \rangle + \langle e_R^2 \rangle + \langle r_R^2 \rangle$$

➤ Error will be overestimated if one attributes all errors to the RS system

➤ For double collocation-based calibration:

$\langle \sigma_B^2 \rangle / (\langle \sigma_R^2 \rangle + \langle e_R^2 \rangle + \langle r_R^2 \rangle)$ is needed

If there is a way to isolate buoy error, double collocation can be a good method.

Double & Triple Collocation

Triple Collocation

$$B = T + \sigma_B + e_B + r_B = T + \sigma_B \quad B: \text{Buoy}$$

$$R = T + \sigma_R + e_R + r_R \quad R: \text{RS}$$

$$M = T + \sigma_M + e_M + r_M \quad M: \text{Model}$$

Truth: 20-min sampled SWH at buoy's locations

Buoy: Reference data, without environment error and representativeness error

$$\langle (B - R)^2 \rangle = \langle \sigma_B^2 \rangle + \langle \sigma_R'^2 \rangle - 2 \langle \sigma_B \sigma_R' \rangle$$

definition

$$\langle (B - M)^2 \rangle = \langle \sigma_B^2 \rangle + \langle \sigma_M'^2 \rangle - 2 \langle \sigma_B \sigma_M' \rangle$$

definition

$$\langle (R - M)^2 \rangle = \langle \sigma_R'^2 \rangle + \langle \sigma_M'^2 \rangle - 2 \langle \sigma_R' \sigma_M' \rangle$$

$$\sigma_R' = \sigma_R + e_R + r_R$$

$$\sigma_M' = \sigma_M + e_M + r_M$$

$$\langle \sigma_R' \sigma_M' \rangle = \langle e_R e_M \rangle + \langle e_R r_M \rangle + \langle r_R e_M \rangle + \langle r_R r_M \rangle$$

relative location and coverage induced errors between RS & Model are not related to environmental variables

Double & Triple Collocation

Triple Collocation

$$\langle \sigma'_R \sigma'_M \rangle = \langle e_R e_M \rangle + \cancel{\langle e_R r_M \rangle} + \cancel{\langle r_R e_M \rangle} + \langle r_M r_R \rangle$$

relative location and coverage induced errors between RS & Model are not related to environmental variables

RS & Model might contain some common signals that are not in buoys $\rightarrow \langle \mathbf{r}_M \mathbf{r}_R \rangle \neq 0$

0. $\langle r_M r_R \rangle$ can be estimated from the difference in variance spectra, but this method could overestimate $\langle r_M r_R \rangle$ because random noises of altimeter are also included spectral difference.

➤ Except for scale difference, representativeness errors from the different spatial-temporal locations can also be a problem

e_M and e_R can be functions of the same variables $\rightarrow \langle \mathbf{e}_M \mathbf{e}_R \rangle \neq 0$.

➤ Linear calibration will also increase $\langle e_M e_R \rangle$ if e_M and e_R are not linear functions of T.

➤ Extremely difficult to estimate $\langle e_M e_R \rangle$

Unless the correlation between σ'_M and σ'_R is known or can be neglected, triple collocation should be also treated with great care!

Estimates of Error Components: Random Error

Assumption 1: $\langle \sigma^2 \rangle$ is stable for a given buoy/altimeter

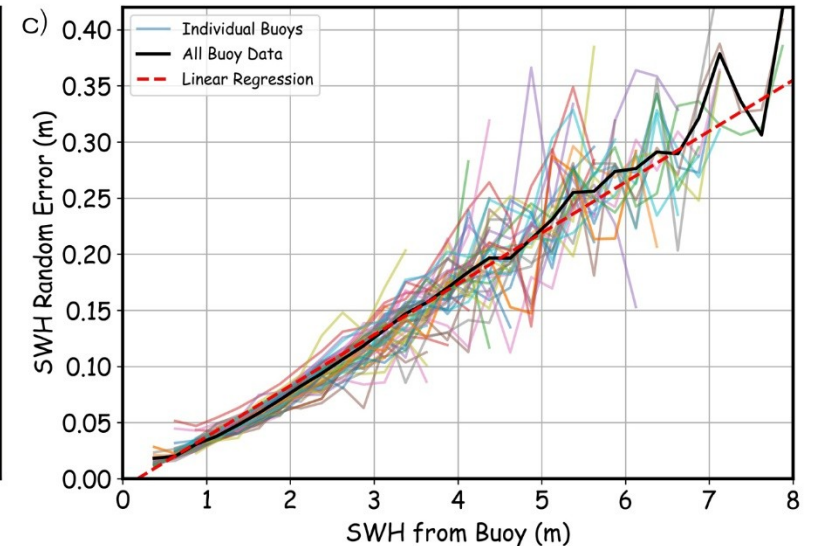
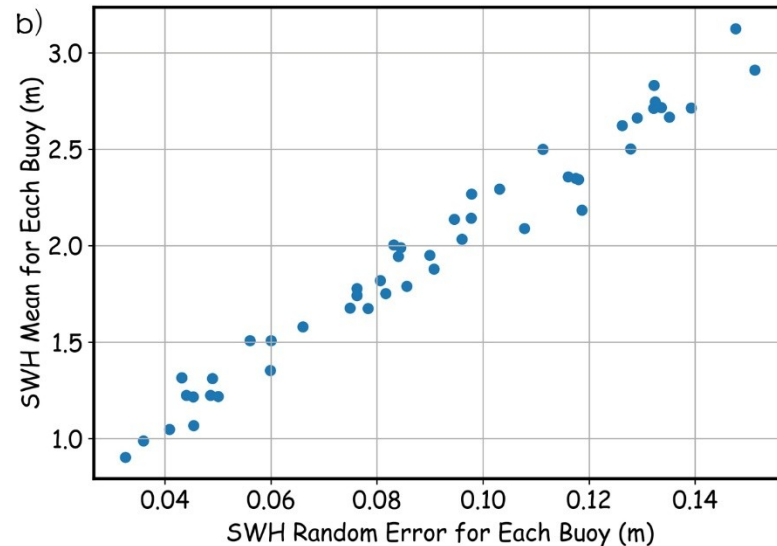
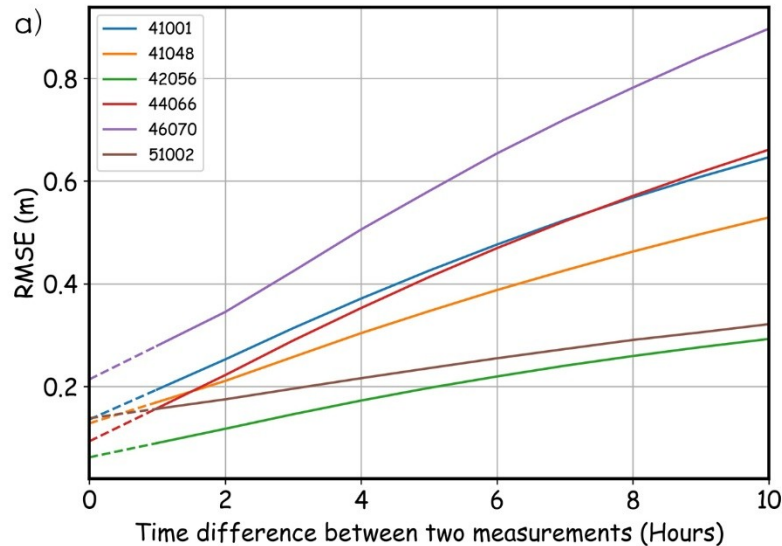
Assumption 2: $\langle r^2 \rangle$ is a function of spatial-temporal distance for the same sensor

For two successive measurements: **“Self-collocation”**
can be used for estimating random error

$$B_n = T + \sigma_{B,n}$$

$$B_{n+i} = T + \sigma_{B,n+i} + r_{B,n+i}$$

$$\langle (B_{n+i} - B_n)^2 \rangle = \langle \sigma_{B,n}^2 \rangle + \langle \sigma_{B,n+i}^2 \rangle + \langle r_{B,n+i}^2 \rangle = 2 \langle \sigma_B^2 \rangle + \langle r_{B,n+i}^2 \rangle \quad \lim_{i \rightarrow 0} \langle (B_{n+i} - B_n)^2 \rangle = 2 \langle \sigma_B^2 \rangle$$

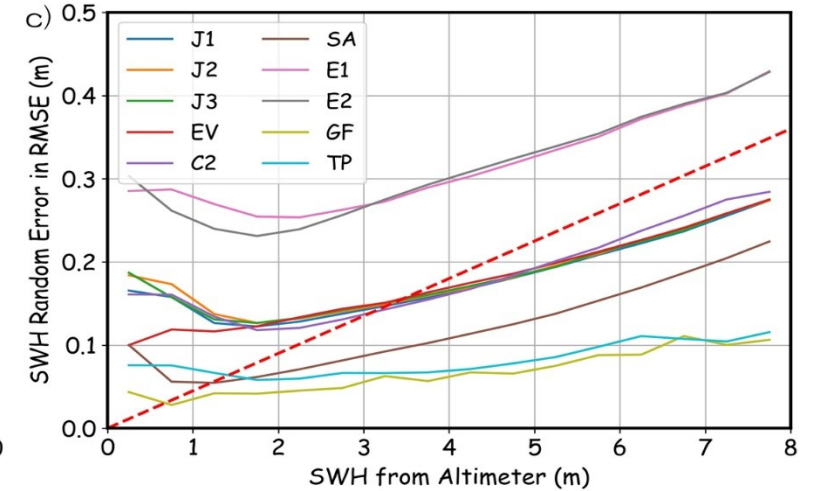
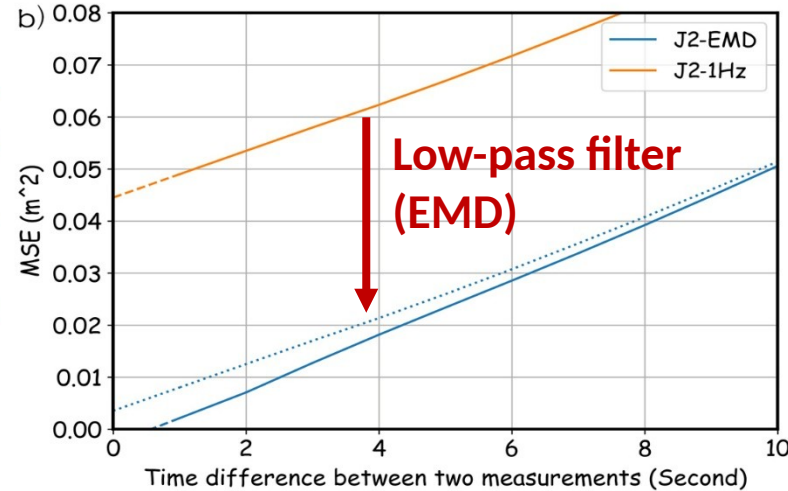
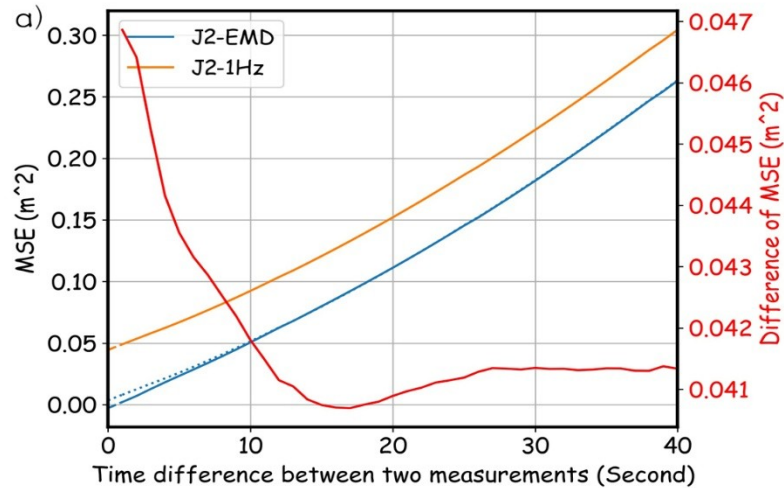


Estimates of Error Components: Random Error

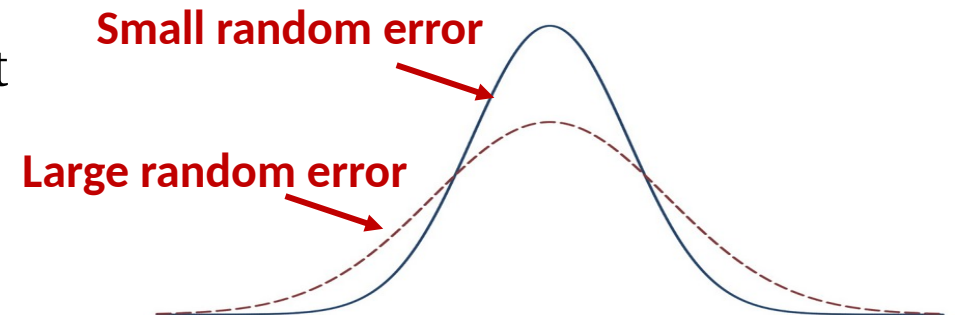
Assumption 1: $\langle \sigma_B^2 \rangle$ is stable for a given buoy/altimeter

Assumption 2: $\langle r^2 \rangle$ is a function of spatial-temporal distance for the same sensor

$$\lim_{i \rightarrow 0} \langle (R_{n+i} - R_n)^2 \rangle = 2 \langle \sigma_R^2 \rangle$$



- Low-pass filter is a good tool to reduce random error, but the representativeness of data is changed.
- Different altimeters have different random errors! (May be important for EVA)

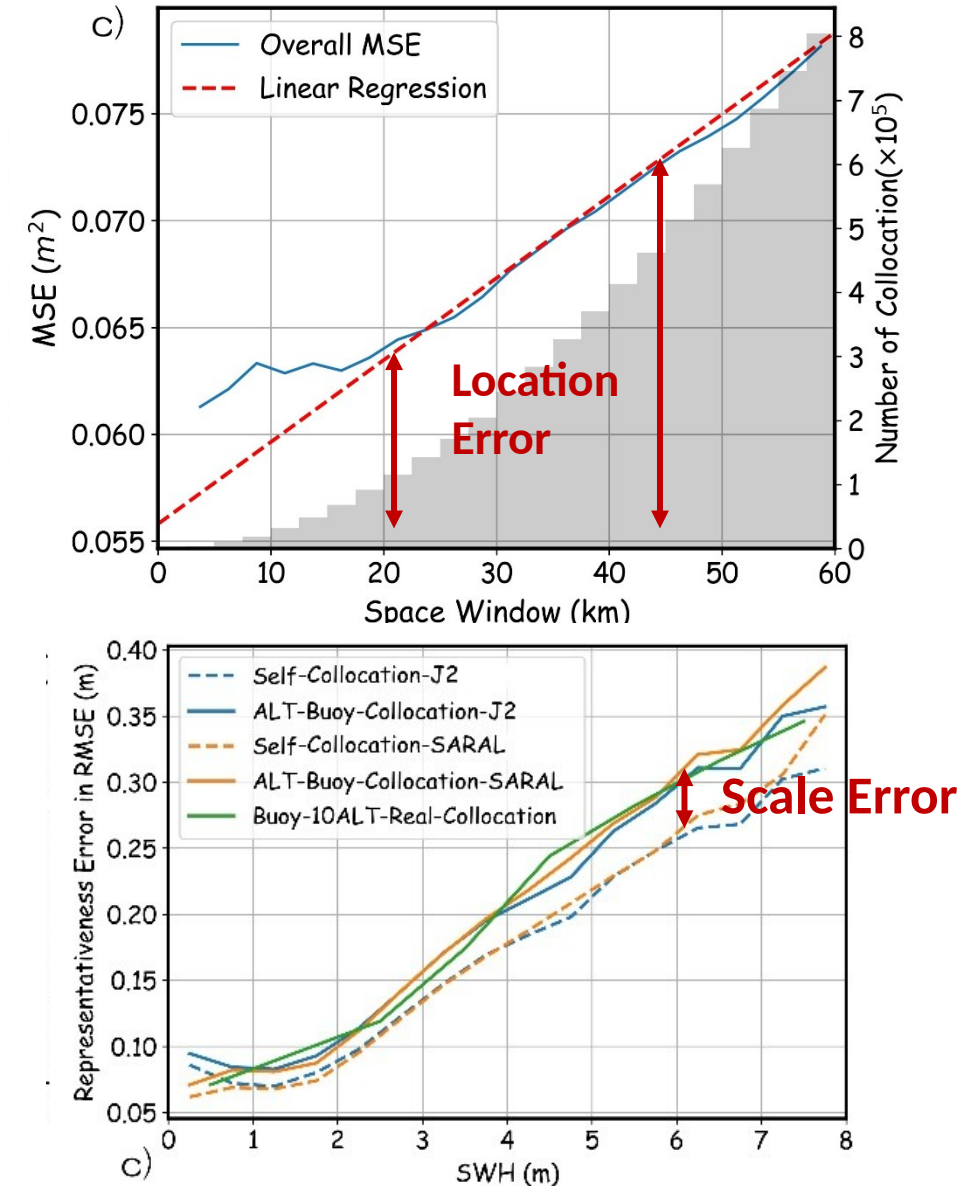


Estimates of Error Components: Representativeness Error

Assumption 3: $\langle r^2 \rangle$ is the superposition of “location” and “scale” errors.

Assumption 4: The magnitude of location errors can be regarded as a linear function of spatial distance (collocation window), and both errors are function of H_s .

- Collocate buoy and RS observations using different windows → Location Errors
- RS self-collocation v.s. RS-Buoy collocation → Scale Errors
- For altimeter v.s. buoy, scale errors seems to be much smaller than location errors.



Estimates of Error Components: Representativeness Error

Reduction of Representativeness Error — Dynamic Collocation

➤ Using a **numerical model** to link satellite and buoy measurements

- Using the difference between locations of RS and buoy in the model to compensate for the representativeness error. Simplest form as:

$$O_{RS} \text{ V.S. } O_{Buoy} + (M_{RS} - M_{Buoy})$$

- Utilizing the spatial correlation of the buoy network, use objective analysis (OA) to correct the model
- Using the corrected model (near buoys) as the reference for RS data
- If there is only one buoy within a large region, the OA is equivalent to direct difference method

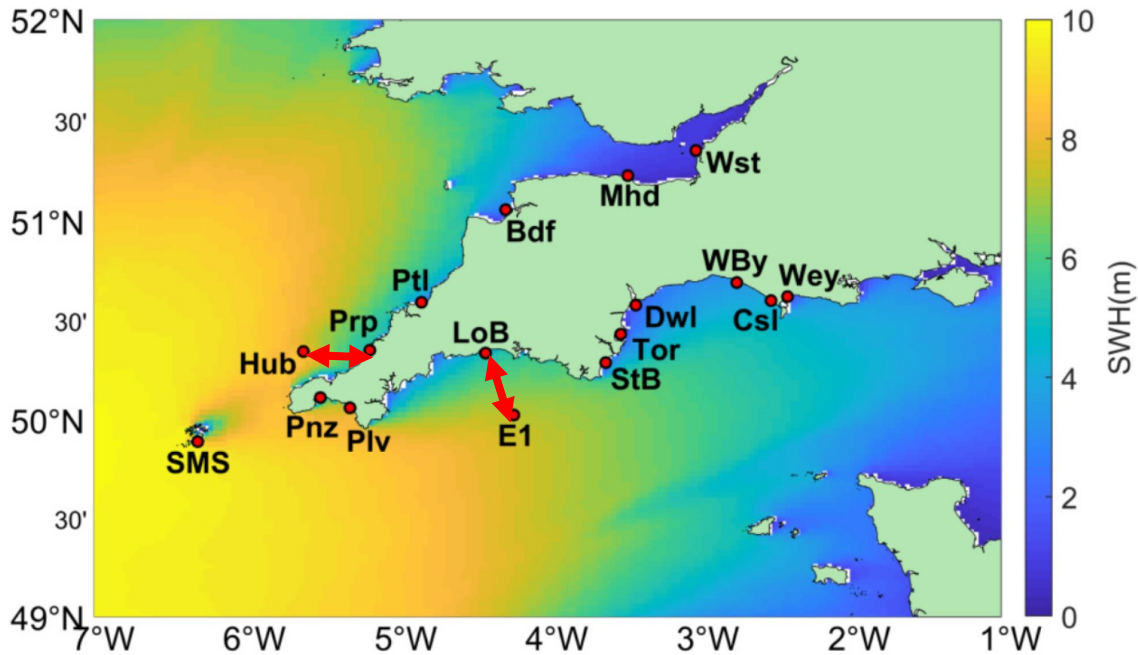
$$F_{Cref}(\phi, \lambda, t) = F_b(\phi, \lambda, t) + \sum_{i=1}^{N_{ref}} w_i [O_i(t) - F_i(t)]$$
$$w_i = \exp[-d_i^2(\phi, \lambda) / 2R^2] / \sum_{i=1}^{N_{ref}} \exp[-d_i^2(\phi, \lambda) / 2R^2]$$
$$R = \min[d_i(\phi, \lambda)] \quad (i = 1, 2, 3, \dots, N_{ref})$$


$$O_{RS} \text{ V.S. } (O_{Buoy} - M_{Buoy} + M_{RS})$$

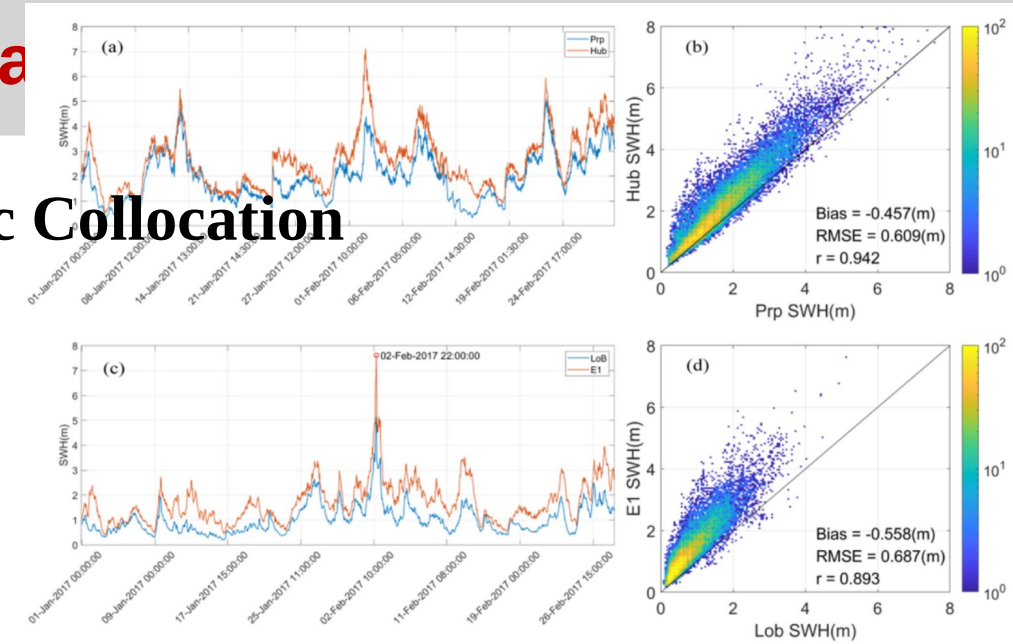
Estimates of Error Components: Representa

Reduction of Representativeness Error — Dynamic Collocation

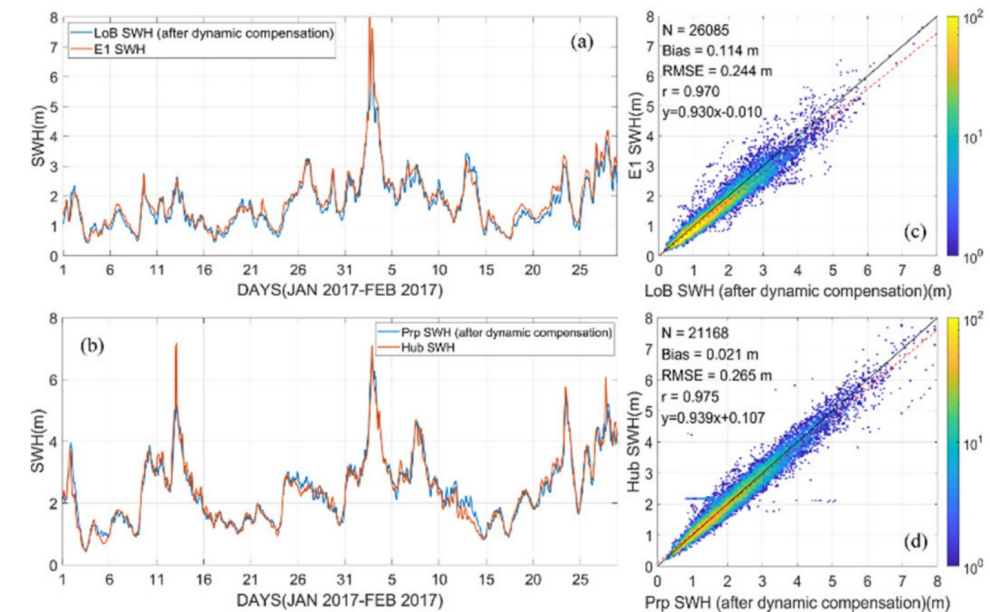
- Complex coastal waves make static windows useless



- Using high-resolution wave hindcast to dynamically collocate two buoys
- The results show the effectiveness of dynamic collocation in coastal regions



Direct comparison between two buoys



Using dynamic collocations

Estimates of Error Components: Environmental Errors

$$\langle (B - R)^2 \rangle$$

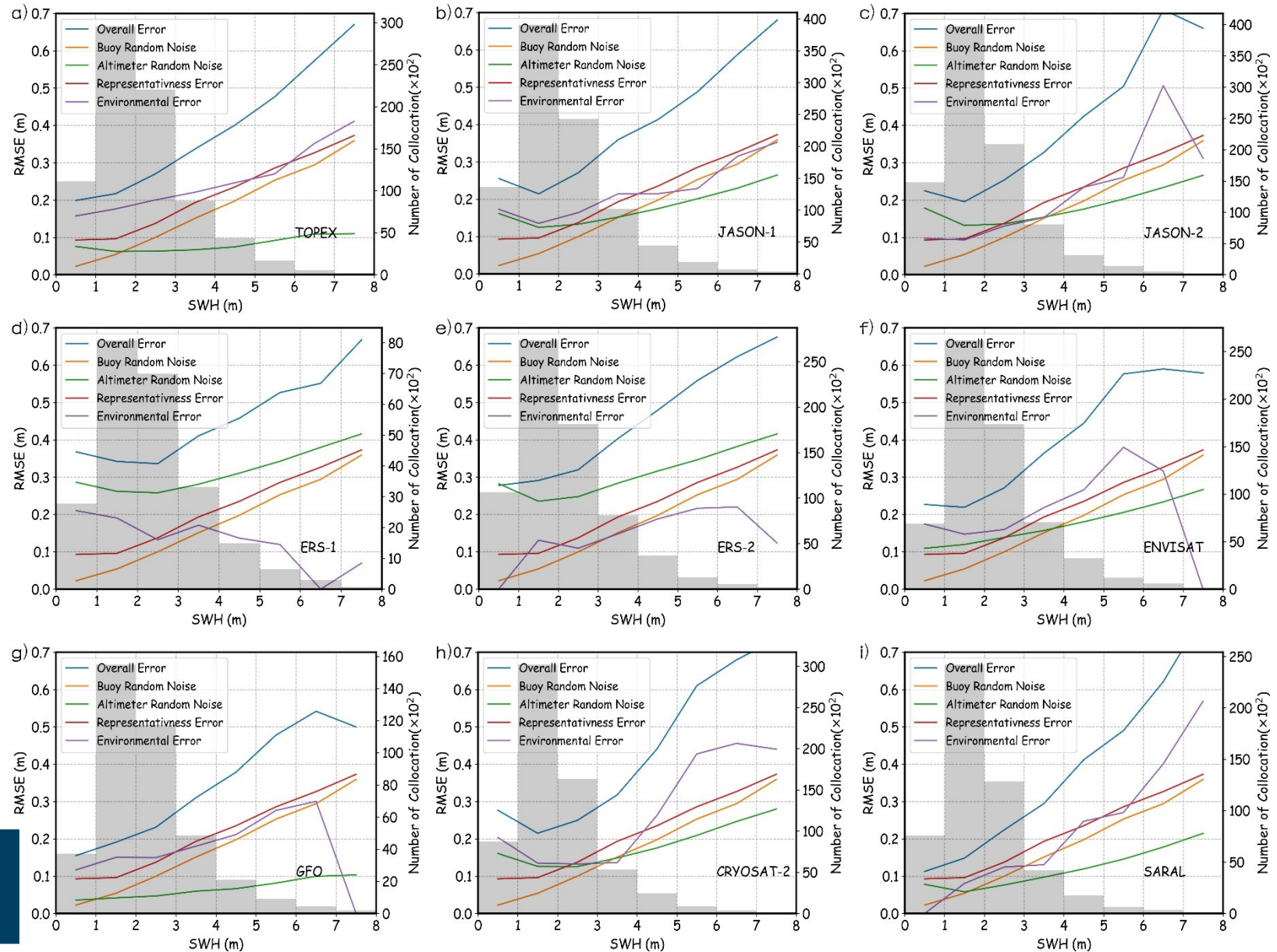
Solved

$$= \langle \sigma_B^2 \rangle + \langle \sigma_R^2 \rangle + \langle r_R^2 \rangle + \langle e_R^2 \rangle$$

The rest: environmental errors

- The three errors are in similar level for most altimeters
- For $SWH > 4$ m the sample might be insufficient for some altimeters

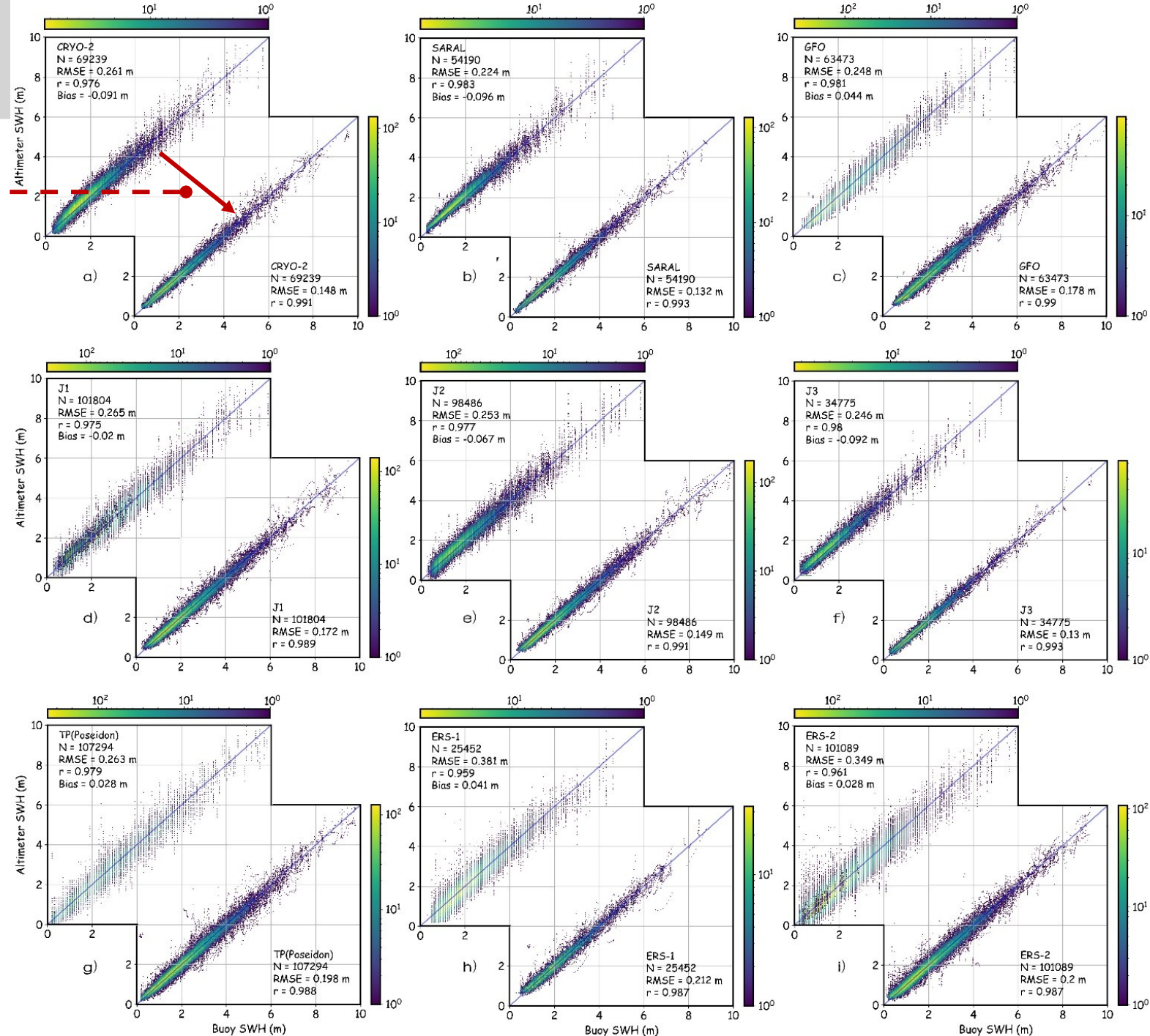
Different error components for different altimeters



Error Reduction

Altimeter Denoising
Buoy Denoising
Dynamic Collocation
ANN Correction

RMSEs between buoy
and altimeter SWHs are
reduced by
25% (TOPEX) ~ 47%
(JASON-3)



A Global SWH Dataset Fused from Multi-Mission Altimetry and WAVEWATCH III Hindcasts

Haoyu Jiang (Haoyujiang@szu.edu.cn)

Department of Oceanography, Shenzhen University, Shenzhen, China

Motivation

➤ Main data sources of wave climate studies:

NWM — **PROS:** “seamless”, consistent

CONS: model is model, not THAT accurate

Buoys — **PROS:** accurate? “direct” measurement

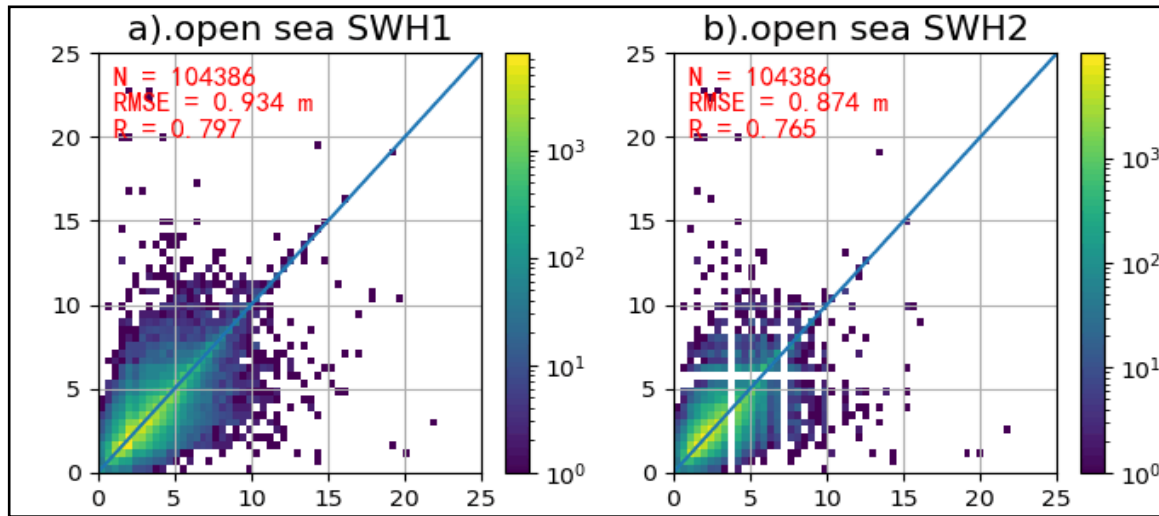
CONS: Sparse, inconsistent with time

ALTs — **PROS:** accurate, consistent (CCI/ Prof. Young)

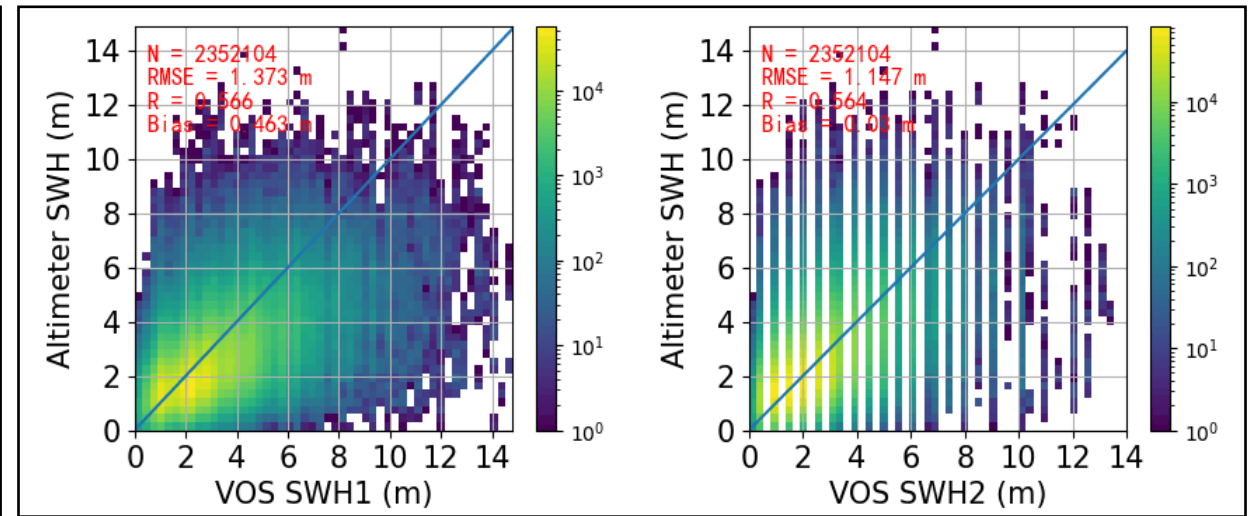
CONS: Sparse (better than buoys)

VOS — **PROS:** long history

CONS: bad quality (sorry to say that)



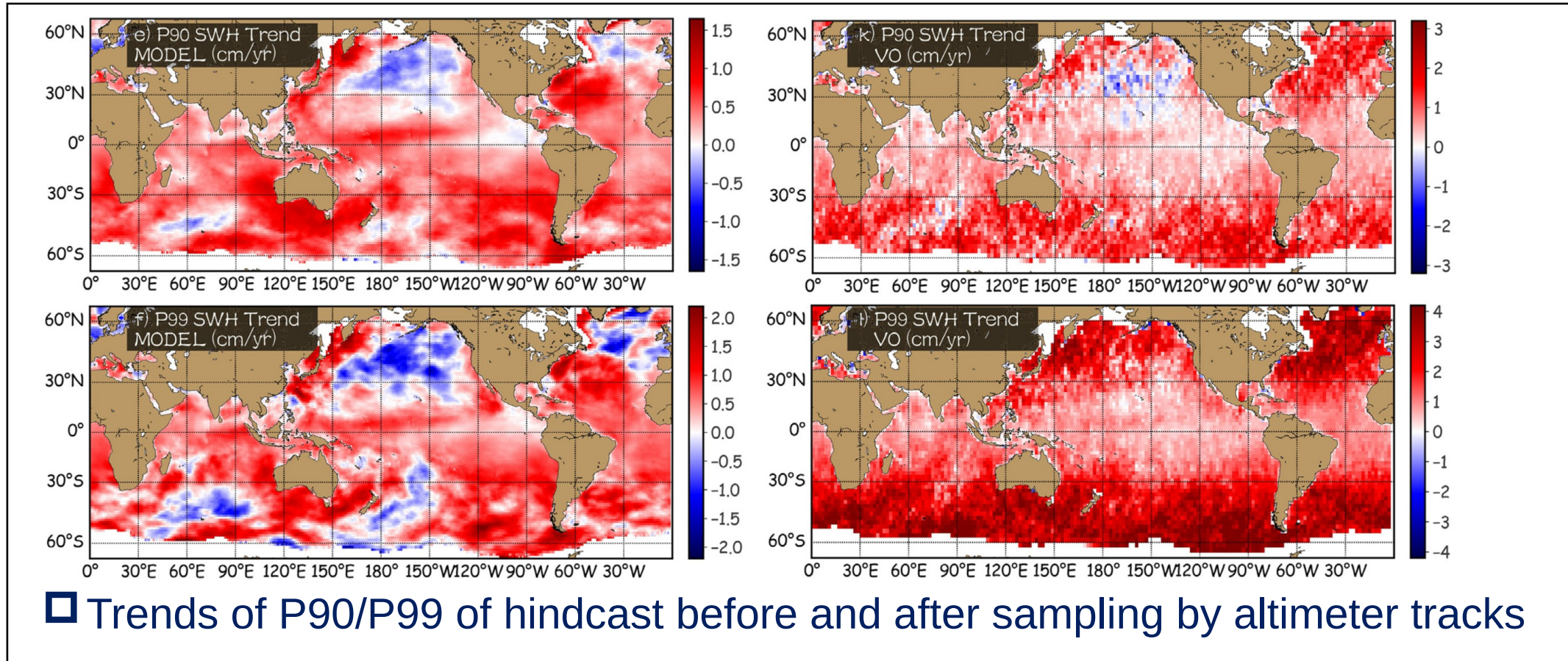
VOS vs VOS



VOS vs ALTs

Motivation

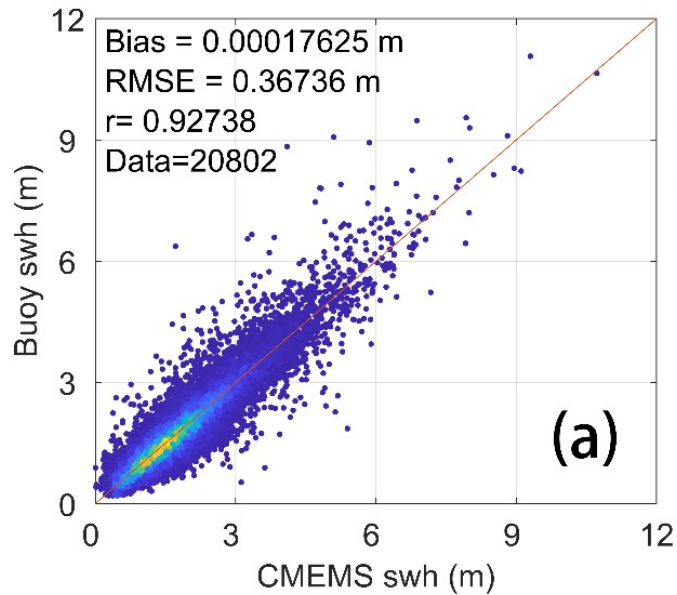
- Sparse sampling can be an issue for wave climate, especially for extreme values



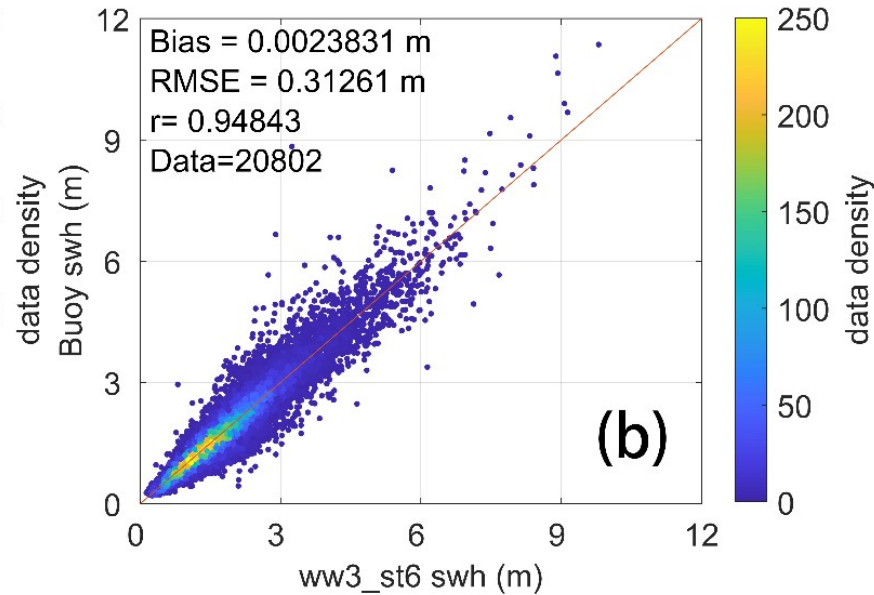
- An “seamless ”altimeter SWH dataset with low error/uncertainty should be useful and convenient

Motivation

➤ Direct Interpolation ? (like AVISO/CMEMS L4 gridded dataset)



Interpolation v.s Buoys



WW3 v.s. Buoys

With 4 satellite, the direct interpolation results are still much worse than WW3

➤ Assimilation? (ERA5/6) Good, but....

1. Difficult for me to run a model 😂
2. Difficult to run WW3 reversely or using N-D Var
 - Future observations cannot correct historical outputs (Am I right about this?)

Methods

- Simple post-fusion might be a good option, because WW3 can provide good hindcasts
- Method for post-fusion : Objective Analysis

$$A(i, j, t) = M(i, j, t) + \sum_{k=1}^N w(i, j, t)_k g(O_k - M_k)$$

Reducing the impact of random error
(eliminate satellite tracks)

$$w(i, j, t)_k = \exp \left[-d_k^2(i, j, t) / 2R(i, j, t)^2 \right] / \sum_{k=1}^N \exp \left[-d_k^2(i, j, t) / 2R(i, j, t)^2 \right]$$

Including temporal interpolation

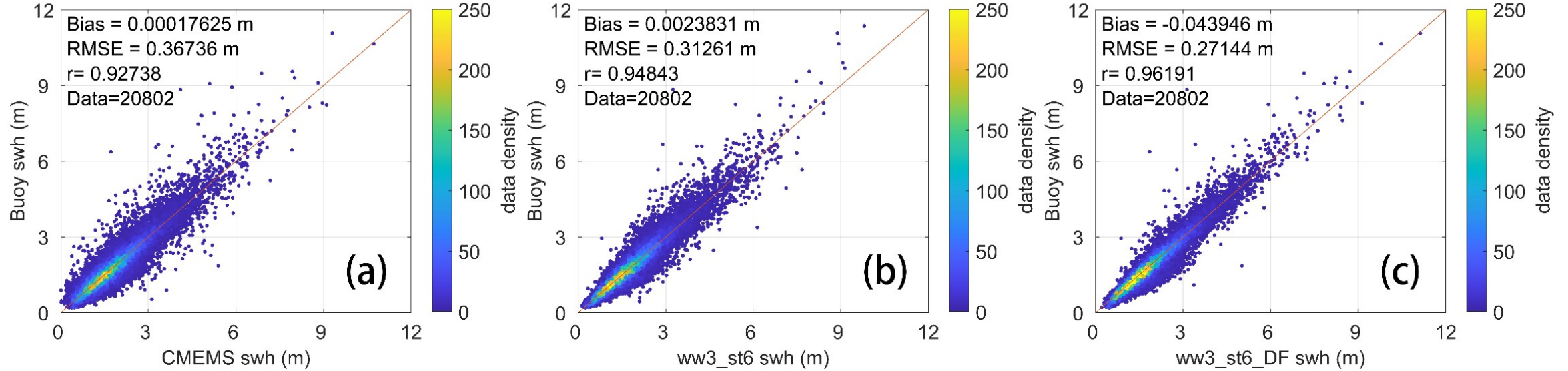
$$R(i, j, t) = \min [d_k(i, j, t)] \quad (k=1, 2, 3, \dots, N)$$

Tricks solving the difference in peak
positions between ALT & Model

$$d_k(i, j, t) = C + \sqrt{(S_k / S_1)^2 + (T_k / T_1)^2 + Q \left[(M_k - M(i, j, t)) / M_k \right]^2}$$
$$d_k(i, j, t) = +\infty \quad \begin{array}{l} (S_k < S_{thr} \ \& \ T_k < T_{thr} \ \& \ r_{i,j,k} > r_{thr}) \\ (S_k > S_{thr} \ \text{or} \ T_k > T_{thr} \ \text{or} \ r_{i,j,k} < r_{thr}) \end{array}$$

- Two versions of the dataset (following AVISO MSLA's idea) :
 - “two-sat” : for climate analysis [66° orbit one + near-polar orbit one]
 - “all-sat” : for other uses [e.g., training AI model]

Data Evaluation — Against buoys



Interpolation v.s Buoy

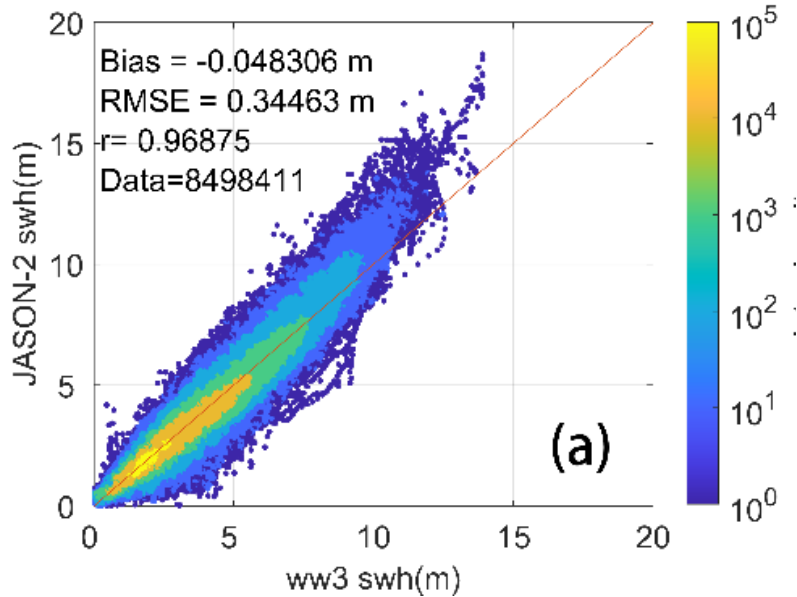
WW3 v.s. Buoy

Fused Dataset v.s. Buoy

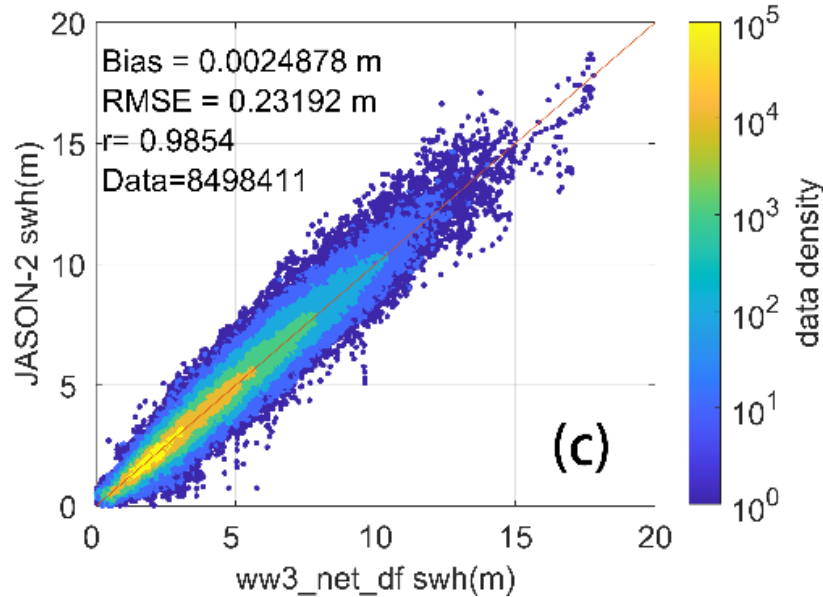
- RMSE and CC are close to the raw 1-Hz altimeter data v.s. buoy

Data Evaluation — Against altimeters

- Using a not-used satellite to evaluate the “two-sat” data



WW3 v.s. Jason-2



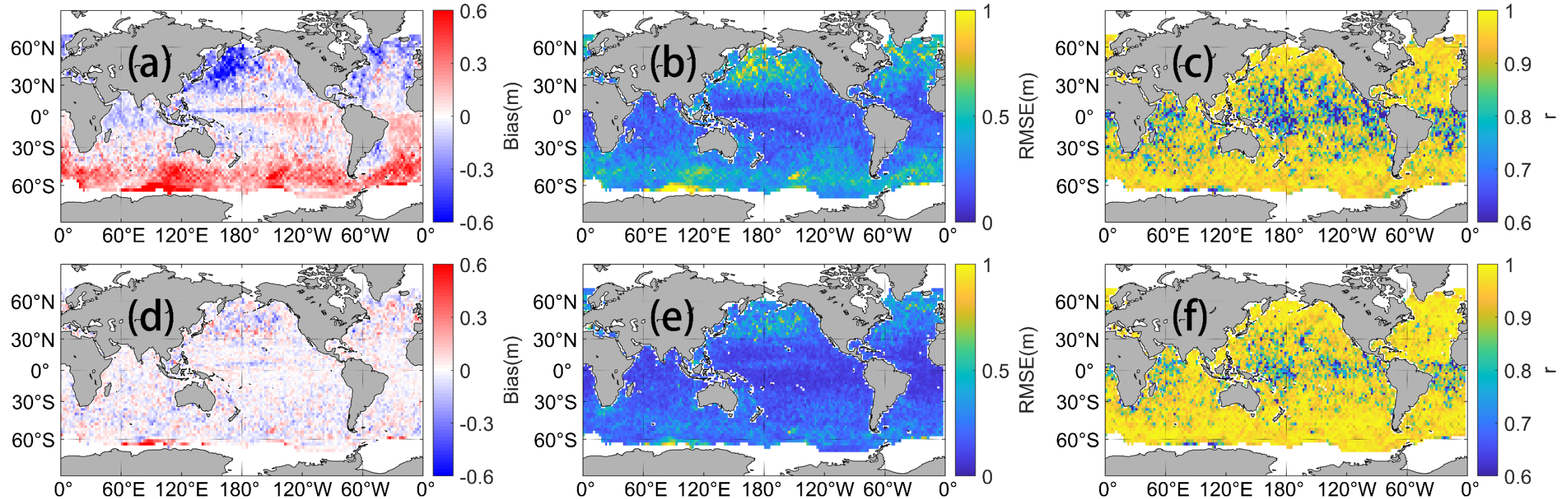
Fused Dataset v.s. Jason-2

RMSE 0.23 ~ 0.25 m for
different combinations

Reference	WW3-ST6	CRYOSAT-2 & JASON-1	CRYOSAT-2 & JASON-2	CRYOSAT-2 & ENVISAT	JASON-1 & JASON-2	JASON-1 & ENVISAT	JASON-2 & ENVISAT
CRYOSAT-2	0.37 m	—	—	—	0.23 m	0.24 m	0.24 m
JASON-1	0.39 m	—	0.25 m	0.24 m	—	—	0.24 m
JASON-2	0.39 m	0.25 m	—	0.24 m	—	0.24 m	—
ENVISAT	0.37 m	0.24 m	0.24 m	—	0.23 m	—	—

Data Evaluation — Against altimeters

- Using a not-used satellite to evaluate the “two-sat” data



WW3 before and after fusing altimeter data

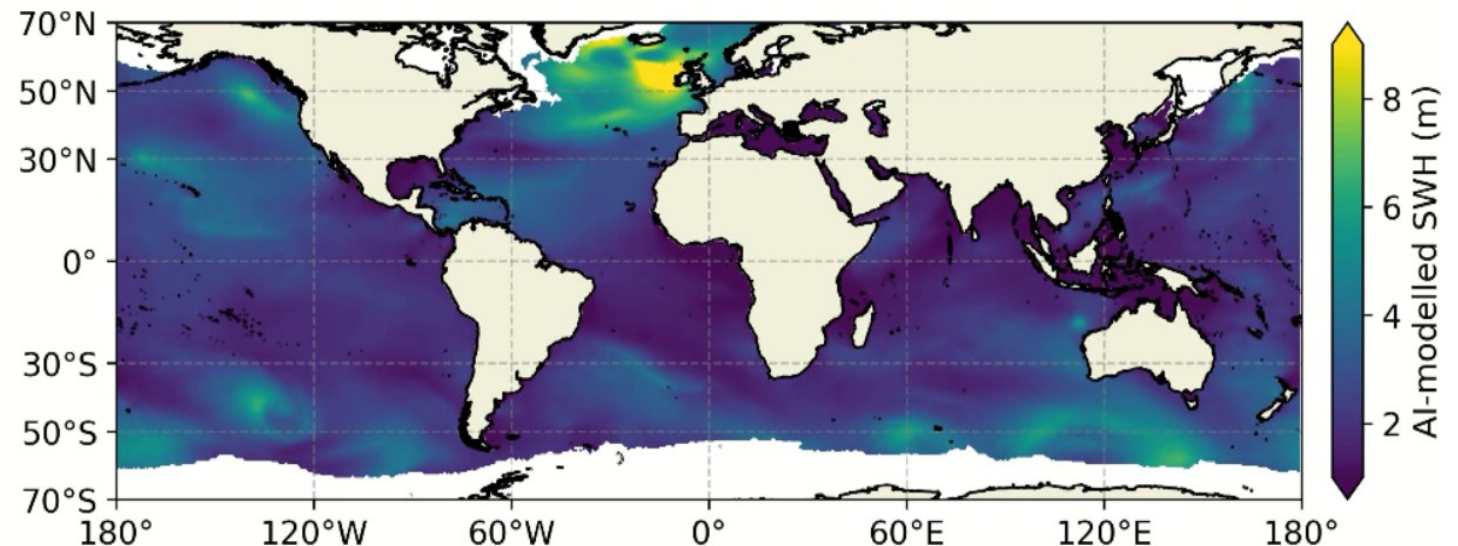
Data Potential Application

- Might be a good dataset for wave climate research
- Due to its high accuracy, it can also be used to train AI wave models

If global historical (~10 days) & current winds are known, the spatial distributions of any wave parameters are almost determined at the current moment [Omitting ocean currents and sea ices].

INPUT : Global wind for the past 10 days → **AI** → **OUTPUT** : Global SWH

- (One of) The most important wave parameter
- Global high-quality satellite observations
→ Good training targets



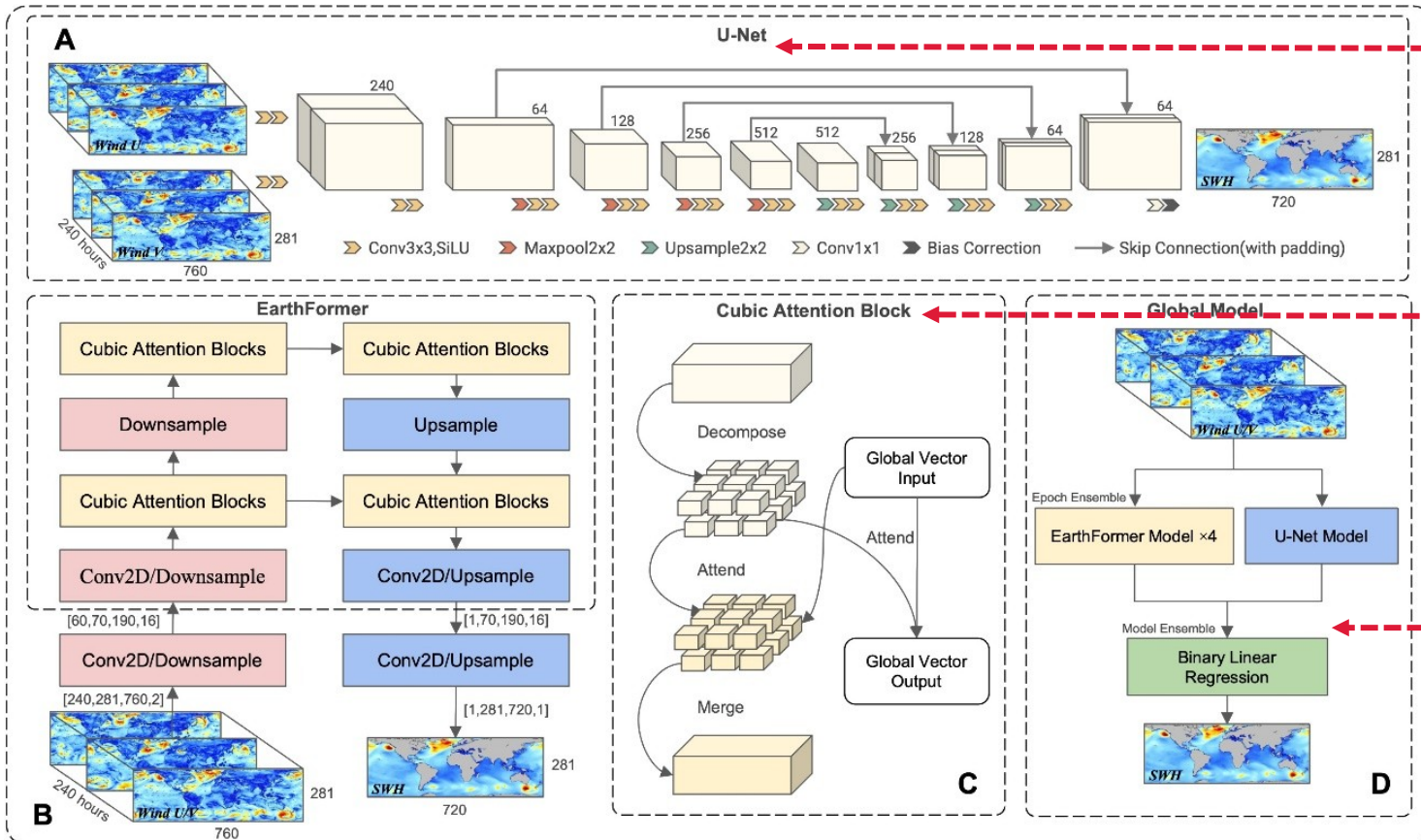
A simple demonstration of AI model training

Pre-training

INPUT : ERA5 winds (U & V) of 10 days

OUTPUT : ERA5 SWH

➤ 2000-2017: Train | 2022: Validation | 1997-1999 & 2018-2020: Testing

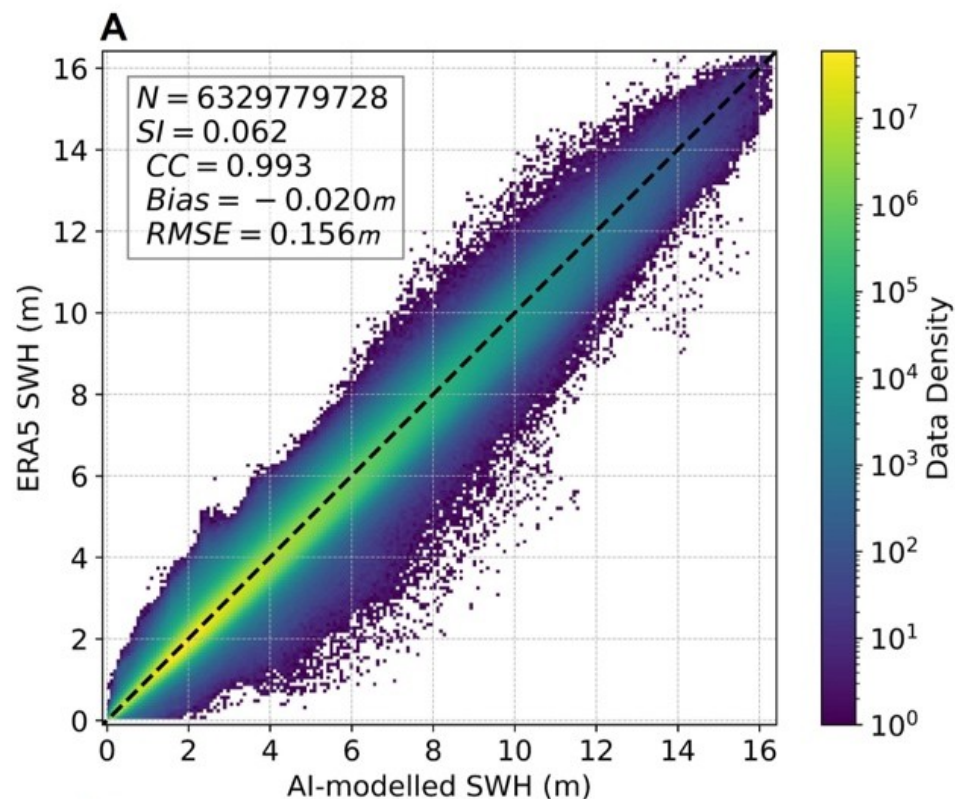


Unet: Local attention by convolutional module to capture local correlation

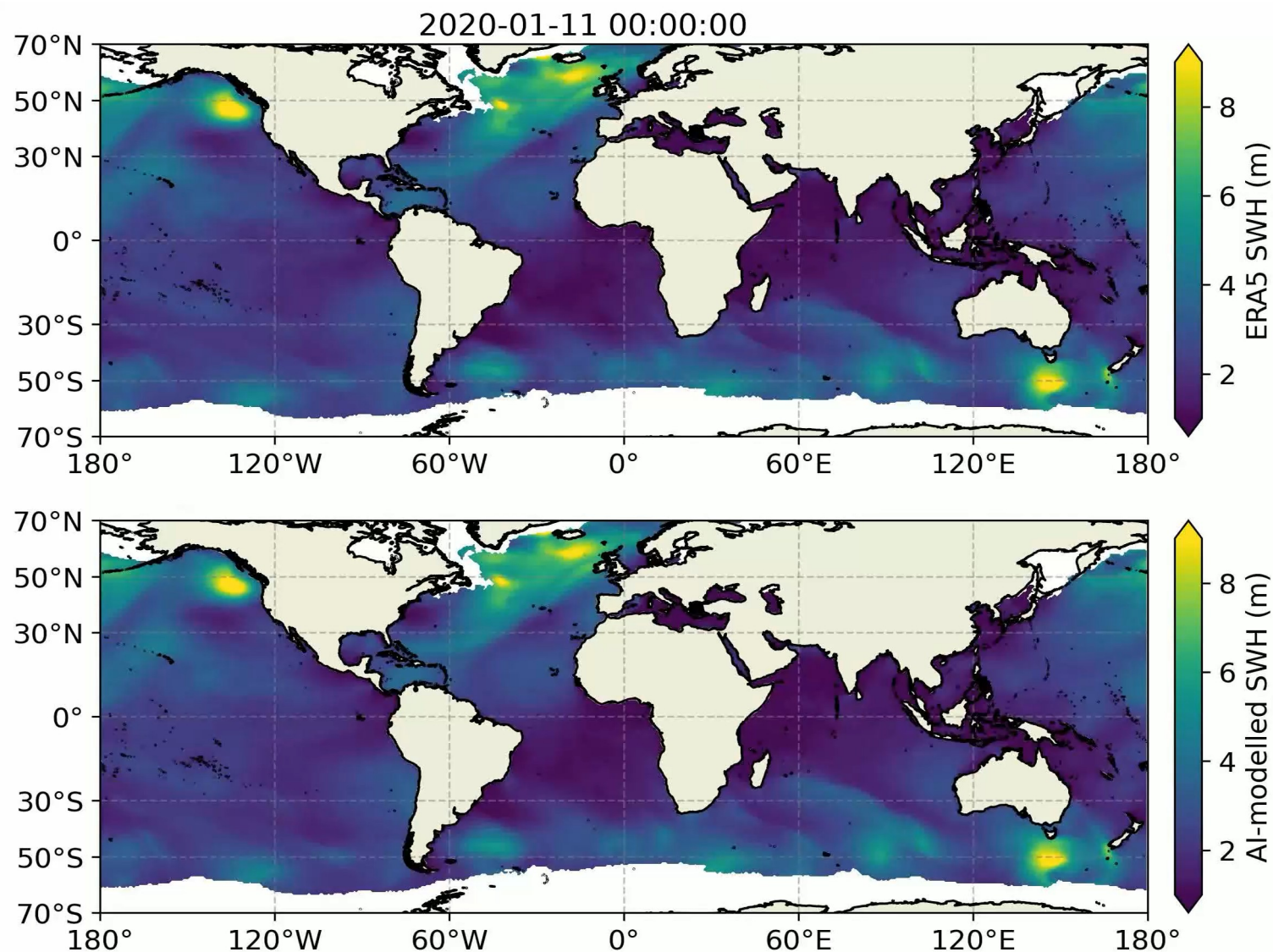
Earthformer: Global attention for tele-connection between remote winds and swells

Model ensemble to obtain better results

Results for the pre-trained model: V.S. ERA5 SWH test set

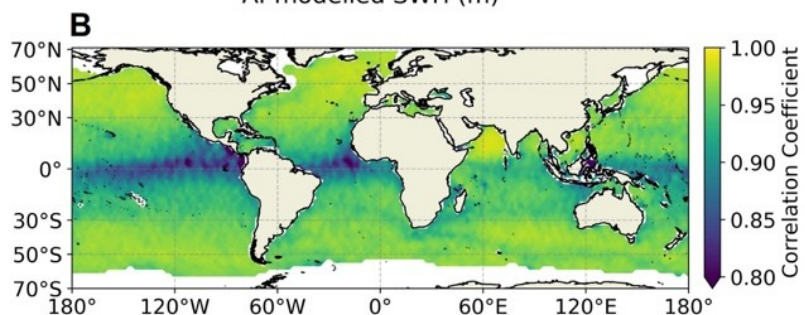
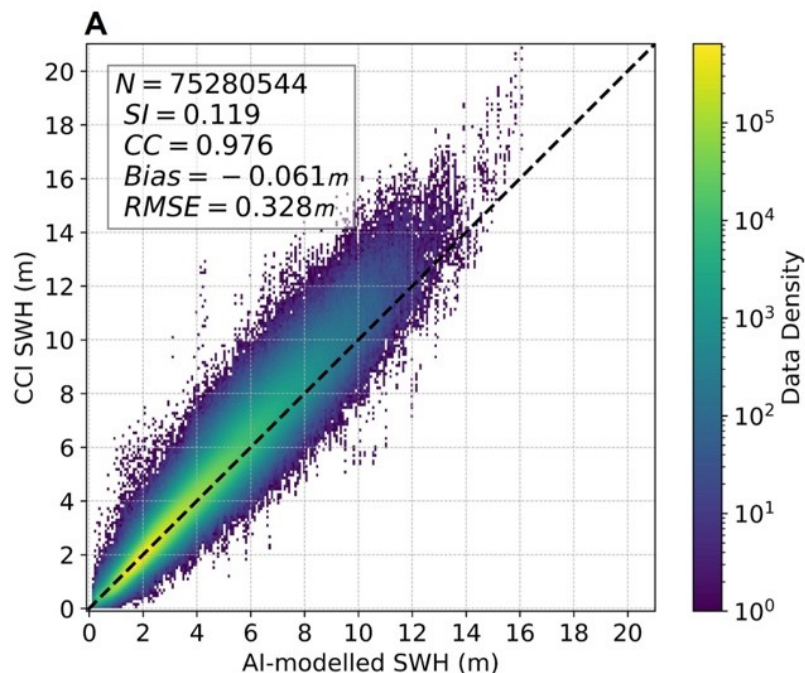


➤ Good agreement with ERA5 SWH

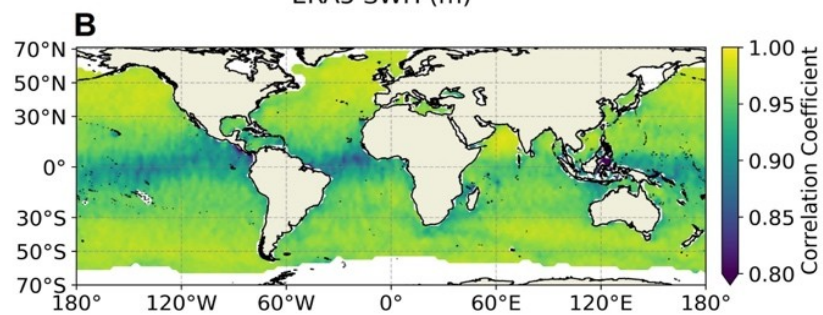
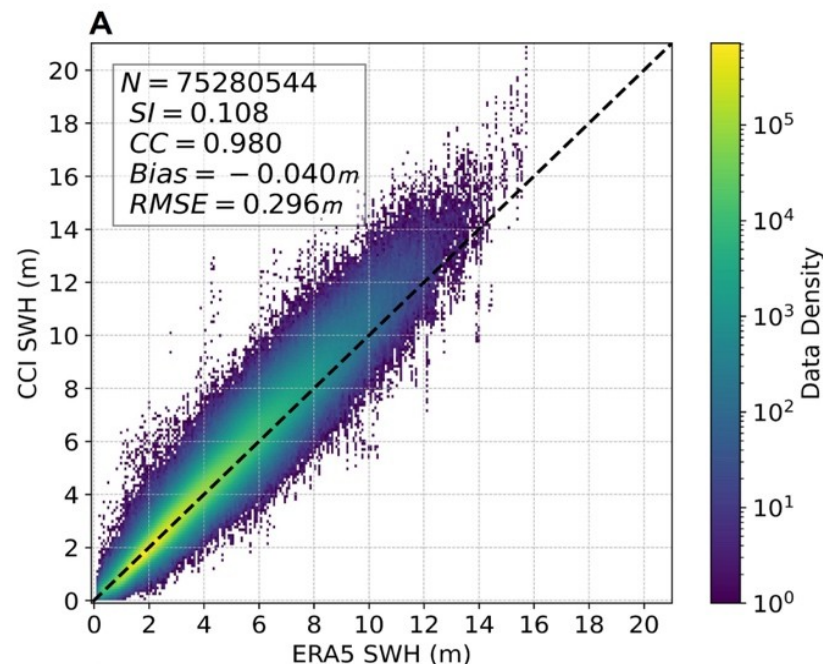


Results for the pre-trained model: V.S. Altimeter Data

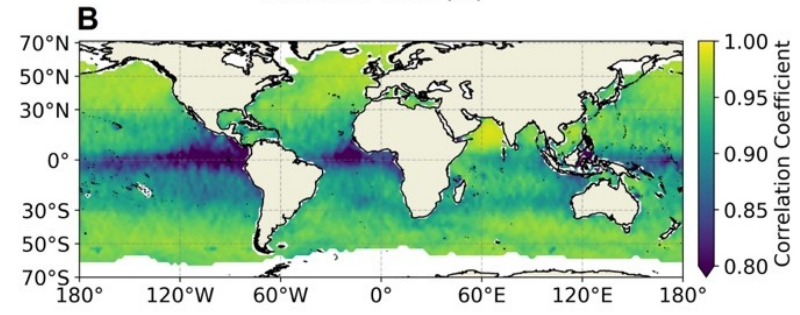
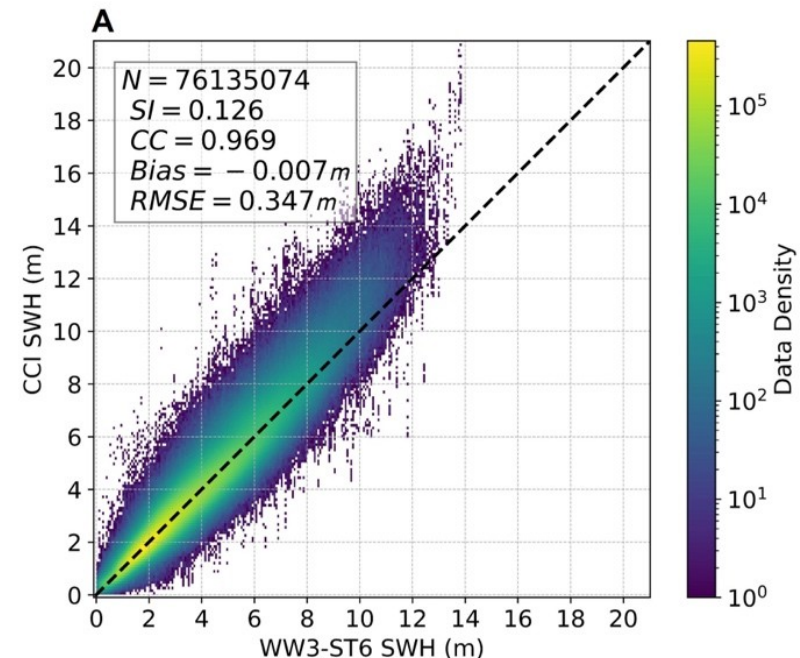
- Only Topex & Sentinel-3A are used for evaluation as they are not assimilated into ERA5



AI vs Altimeter



ERA5 vs Altimeter



WW3-ST6 vs Altimeter

Transfer Learning using Altimeter Data

- The CCI-sea state dataset was merged with the WW3-ST6 data using an objective analysis method from 2016 to 2020 (At least four satellite in orbit during this period) to generate the target dataset.
[2017-2019 for training, 2016 for validation, 2020 for testing]
- During the fine-tuning process, the parameters of the encoder layers of the AI model were fixed, and only the decoder layers were updated through back-propagation.

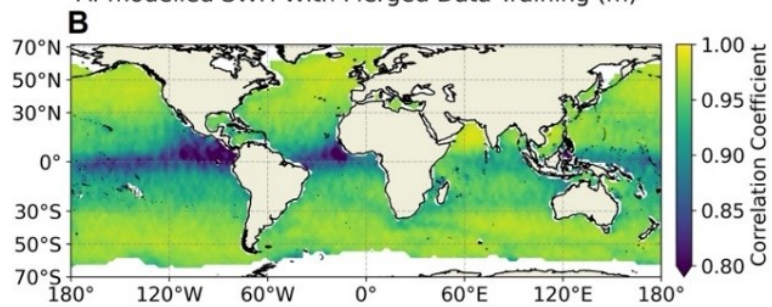
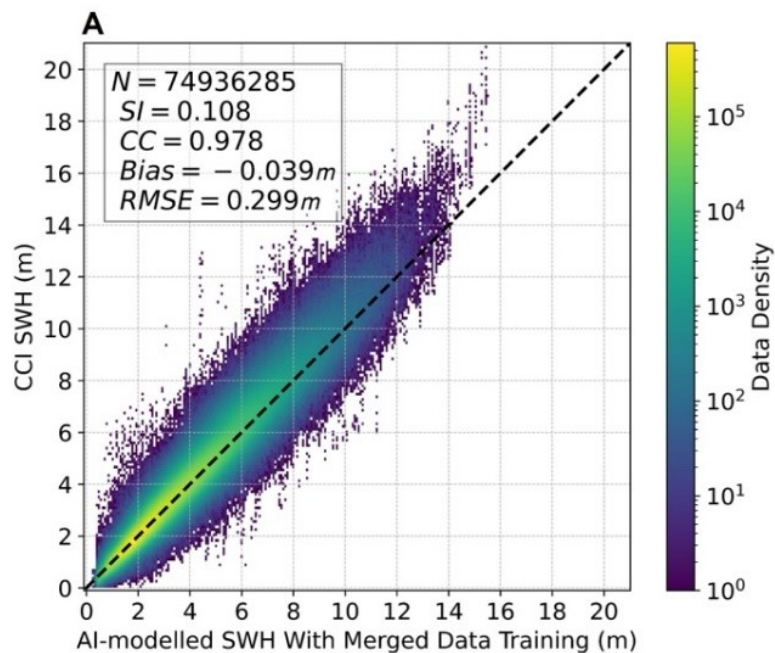
$$Loss = \frac{1}{n} \sum_{t=1}^{Time} \sum_{j=1}^{Lat} \sum_{i=1}^{Lon} \left[W_{i,j,t} (y_{i,j,t} - x_{i,j,t}) \cos \theta_j \right]^2 \quad (S5)$$

$$W_{i,j,t} = 1 / \left[R(i, j, t) + c \right] \quad (S6)$$

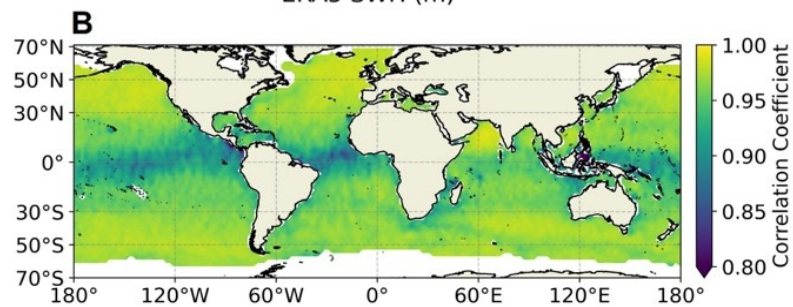
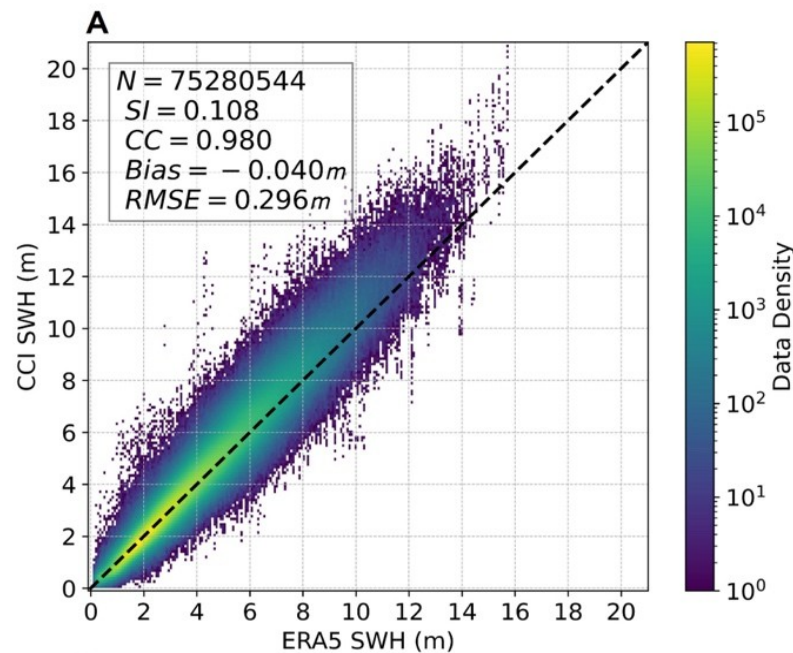
Transfer learning: To addresses the problem of "small" datasets when the target "small" dataset shares similar characteristics with a "large" dataset.

Results for the fine-tuned model: V.S. Altimeter Data

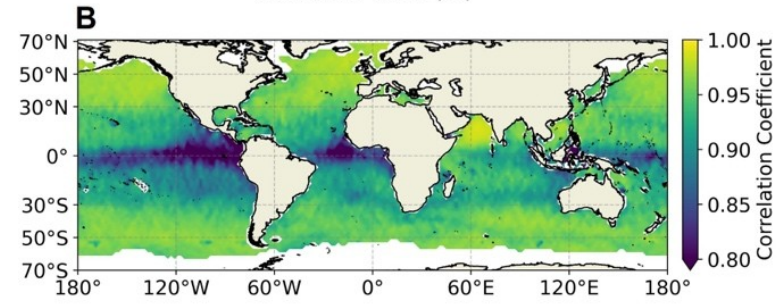
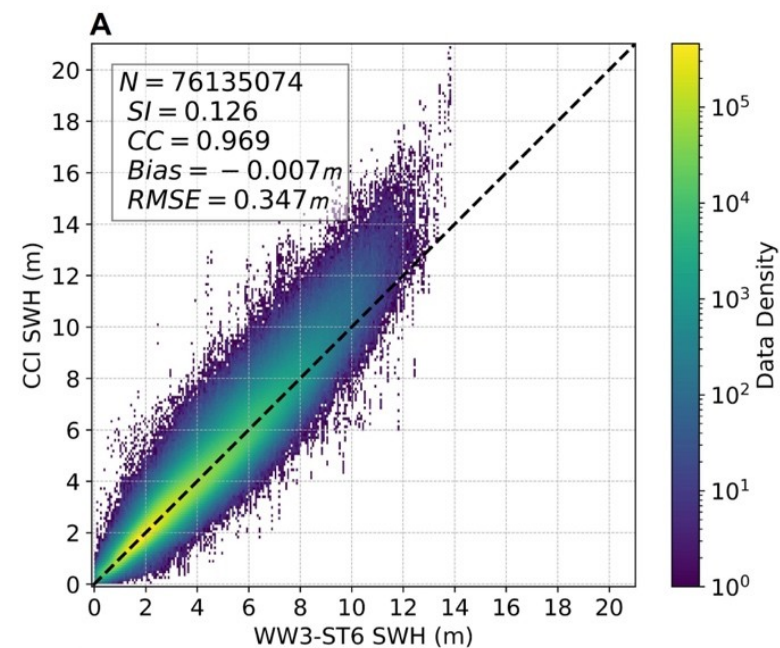
- Only Topex & Sentinel-1A are used for evaluation as they are not assimilated into ERA5



AI vs Altimeter

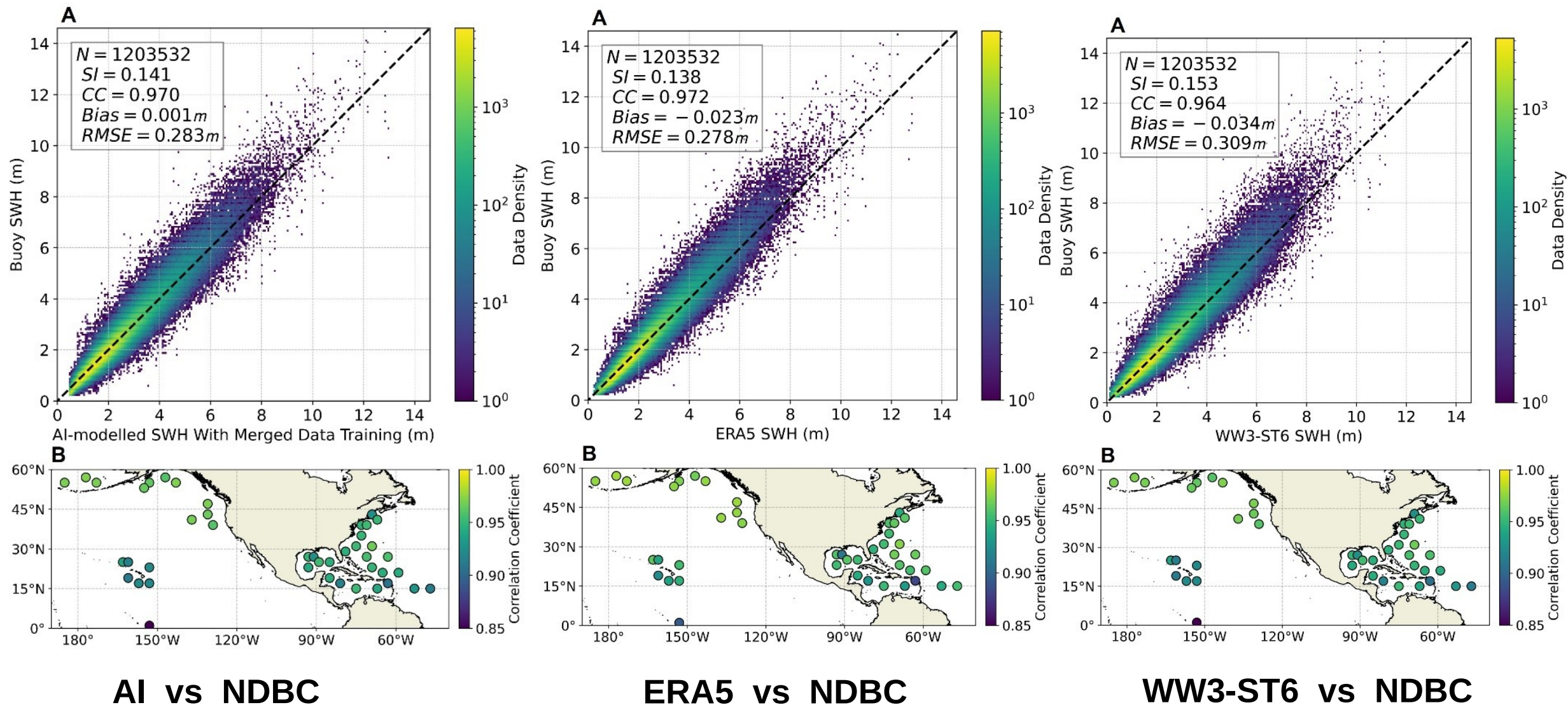


ERA5 vs Altimeter



WW3-ST6 vs Altimeter

Results for the fine-tuned model: V.S. NDBC buoy data



Decompose SWH errors into three components:

- *Random errors* (zero-mean, uncorrelated with external variables)
- *Representativeness errors* (mismatches in location & resolution between “truth” & measurements)
- *Environmental errors* (dependent on external/environmental variables)

Estimation approach & suppression strategies :

- Random → extrapolated from self-collocation results to zero spatial/temporal separation
→ filtering (e.g., running average, EMD)
- Representativeness → MSE between different window size & between RS-RS & RS-Buoy collocations
→ dynamic collocation
- Environmental → residual part, since their MSEs follow the superposition principle
→ recalibration with additional environmental parameters

Results:

- Dominant error components vary across altimeters, though often of similar magnitude
- The method might be applicable to other RS products (wind, currents, SST, ocean color, etc.).

A new dataset fusing altimeter data into WW3 hindcast:

- Seeming to be in good accuracy
- Might be suited for climate research
- A good training target for AI wave models

THANK
YOU!