



Impacts of Measurement Uncertainty (on validation)

Lachlan McKinna, Go2Q

08 June 2022



Overview



Meeting doughnuts

1. Question: can we incorporate uncertainties when computing skill metrics?

Every measurement (satellite & in situ) has uncertainty.

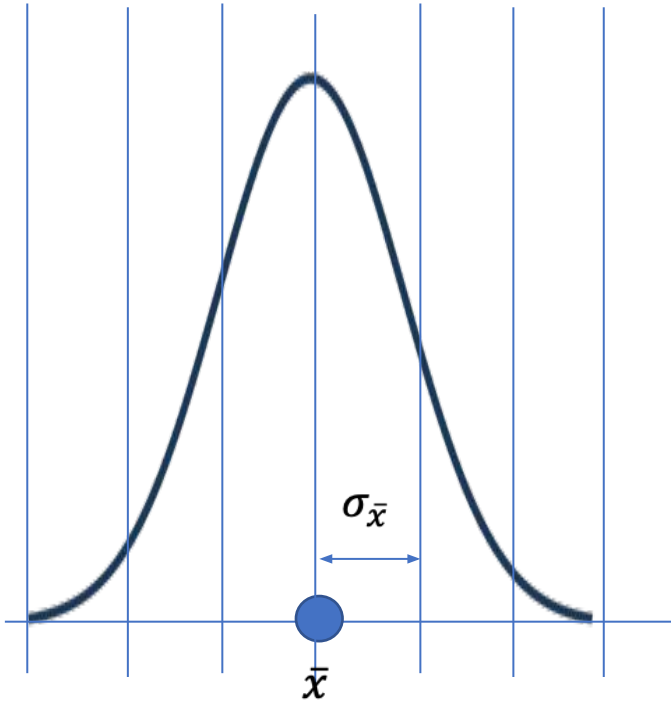
2. Solution: “degree of overlap” metric when computing validation different metrics.

We have found including uncertainties changed validation metrics (mean bias & MAE).

3. Can we use uncertainties to improve the way we visualize uncertainties?

New “zeta score” plots are color-coded and easy to interpret.

Uncertainties in OC data products – why do it?



Quantifying uncertainty in derived ocean color data products (i.e., measurands) is highly valuable, allowing end-users to: assess if datasets are fit-for-purpose, assess if observed temporal change is greater than uncertainty, assimilate uncertainties into climate models, and assess consistency among sensors (Maritorena et al., 2010; Gould et al., 2014). Additionally, a thorough understanding of uncertainty sources within a model may help guide the decisions of scientists when developing new satellite algorithms.

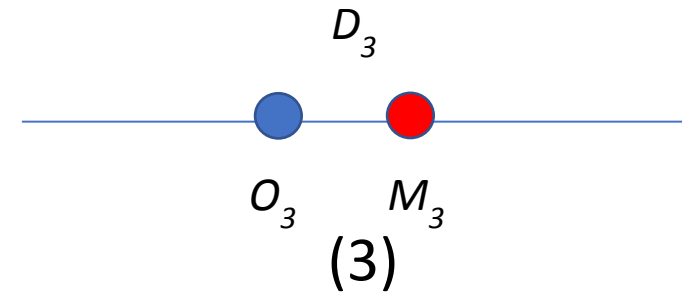
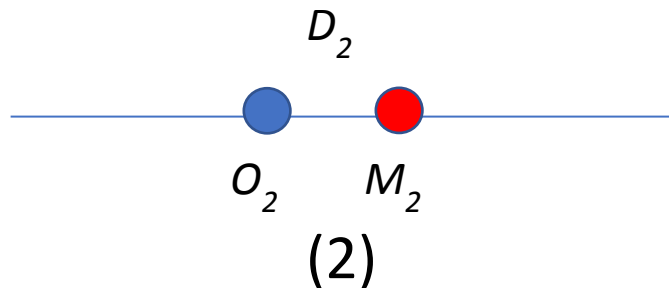
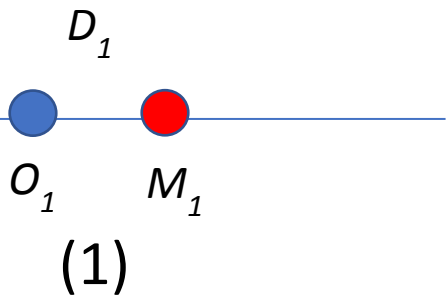
McKinna et al. (2019)

Existing gap – *what about comparing measurements during validation?*

Question:

Based on horizontal distance (D) between the blue and red dots, which pair(s) below would you consider to be different: 1, 2, or 3?

Answer: D is the same for all!

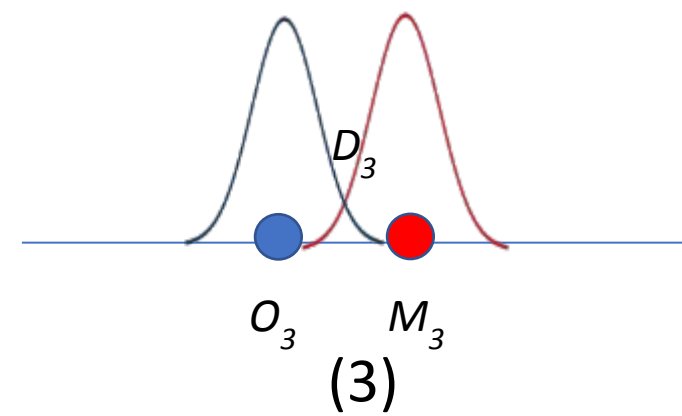
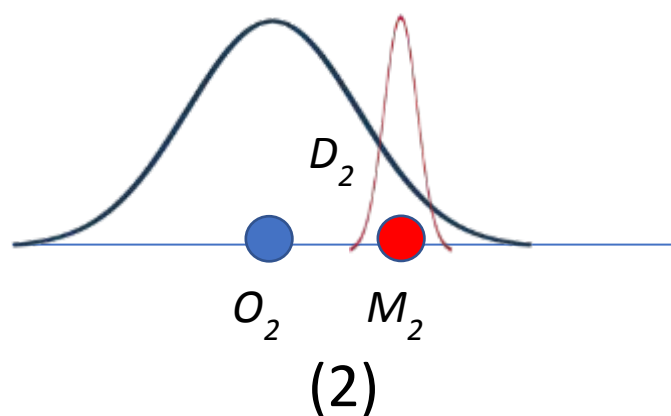
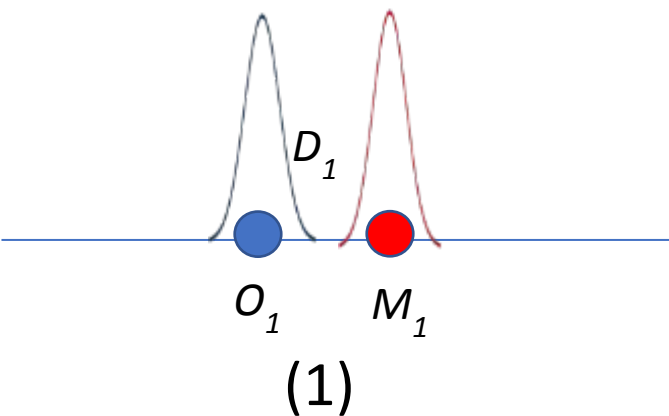


Question:

Based on horizontal distance (D) between the blue and red dots, which pair(s) below would you consider to be different: 1, 2, or 3?

This time, we'll consider measurement uncertainty and draw a probability density function around each point ...

Answer: (a) yes, (b) no, (c) somewhat.

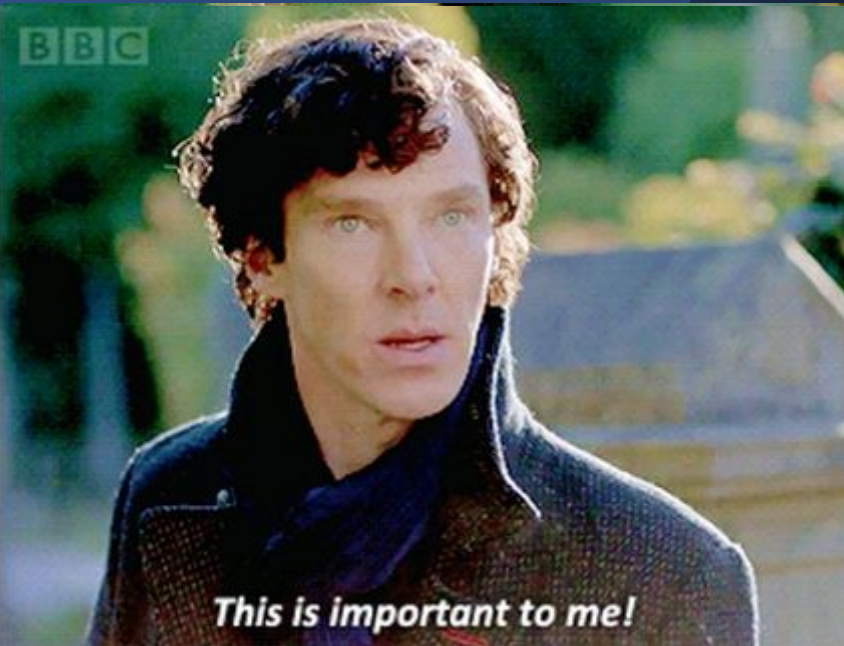


Why is this important for OC validation?

Satellite-derived (M) and the *in situ* (O) measurements are considered exact.

This is NOT true – M & O both have inherent uncertainty.

- We should probably correct validation metrics for measurement uncertainty
- If uncertainties are known, we can improve how we present our results graphically



Validation metrics

$$\text{mean bias} = \frac{1}{N} \sum_{i=0}^N M_i - O_i$$

$$MAE = \frac{1}{N} \sum_{i=0}^N |M_i - O_i|$$

MAE: mean absolute error
See Seegers et al (2018) for more on validation metrics

A method to account for overlapping PDFs

For mean bias and MAE, we compute the difference between the satellite observed (O_i) and in situ measurement (M_i) data pairs:

$$D_i = M_i - O_i$$

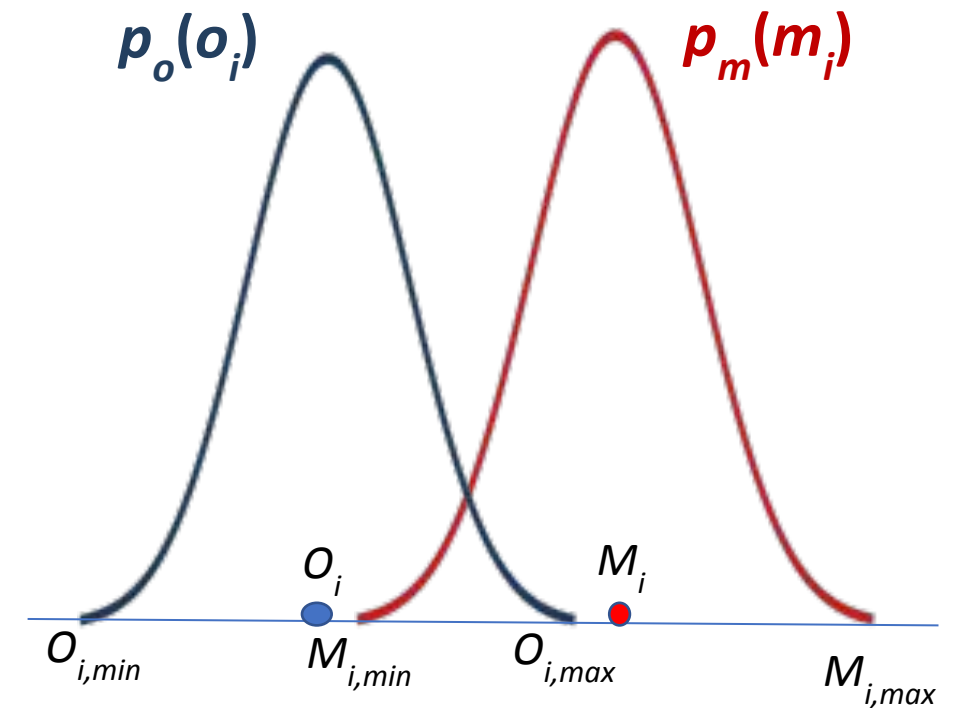
We correct difference with correction factor (CF):

$$CF_i = 1 - DO_i$$

(DO_i) is the *degree of overlap metric* proposed by Harmel et al (2010) (see paper for calculus).

Corrected difference is:

$$D'_i = CF_i D_i$$



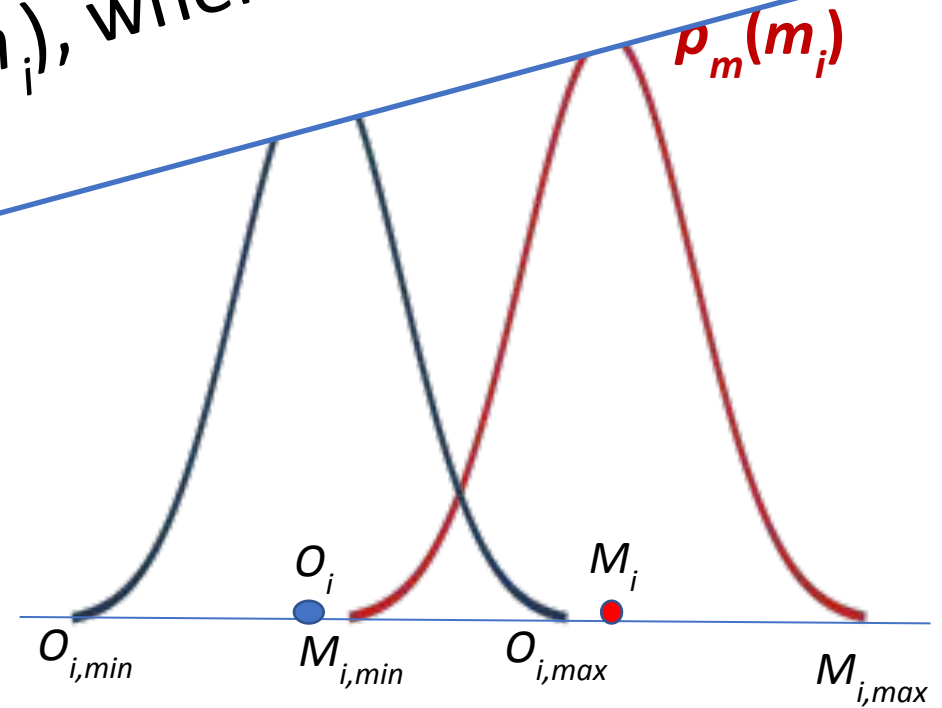
A method to account for overlapping PDFs

For mean bias and MAE, we compute the difference between the satellite observed (O_i) and in situ measurement (M_i) data pairs:

- Less weight is applied when DO_i approaches 0
- For completely overlapping $p_o(o_i)$ and $p_m(m_i)$, where $DO_i = 1$, no difference can be discerned

Corrected difference is:

$$D'_i = CF_i D_i$$



Validation metrics

$$\text{mean bias} = \frac{1}{N} \sum_{i=0}^N M_i - O_i$$

$$MAE = \frac{1}{N} \sum_{i=0}^N |M_i - O_i|$$

Corrected validation metrics

$$\text{mean bias}' = \frac{1}{N} \sum_{i=0}^N CF_i(M_i - O_i)$$

$$MAE' = \frac{1}{N} \sum_{i=0}^N |CF_i(M_i - O_i)|$$

Our study

- We developed an empirical algorithm for derived $b_{bp}(555)$ (not shown here)
- We compared the new model with the GIOP and a Chl-based model (Huot et al. 2008).
- We used the OC-CCI bio-optical dataset (Valente et al. 2015) to validate each model
- 5% relative uncertainties in R_{rs} and $b_{bp}(555)$ were assumed when computing corrected metrics



Did it change the matchup result?

Table 2
Model Difference Statistics Comparing Three Models: LH-Based Model, GIOP, and Huot

$b_{bp}(555)$ range	Model	N	R^{2*}	Slope*	Bias (m^{-1})	Bias' (m^{-1})	MAE (m^{-1})	MAE' (m^{-1})	bias _{log} (unitless)	bias' _{log} (unitless)	MAE _{log} (unitless)	MAE' _{log} (unitless)	No. wins
All data	LH	326	0.730	1.35	3.90×10^{-4}	2.61×10^{-4}	8.01×10^{-4}	3.96×10^{-4}	1.21	1.12	1.33	1.16	0
	GIOP	326	0.733	1.04	1.52×10^{-4}	9.51×10^{-5}	6.75×10^{-4}	3.86×10^{-4}	1.06	1.03	1.27	1.15	10
	Huot	326	0.699	1.40	-7.38×10^{-4}	-5.49×10^{-4}	9.15×10^{-4}	6.59×10^{-4}	0.812	0.834	1.37	1.27	0
$<1.25E-3$ m^{-1}	LH	60	0.225	0.764	6.72×10^{-4}	5.19×10^{-4}	6.75×10^{-4}	5.19×10^{-4}	1.73	1.55	1.73	1.55	0
	GIOP	60	0.049	0.614	2.65×10^{-4}	1.77×10^{-4}	3.81×10^{-4}	2.30×10^{-4}	1.24	1.15	1.48	1.29	0
	Huot	60	0.235	1.01	1.51×10^{-4}	1.47×10^{-4}	1.96×10^{-4}	1.52×10^{-4}	1.17	1.09	1.23	1.10	10
$\geq 1.25E-3$ m^{-1}	LH	266	0.548	1.24	3.25×10^{-4}	2.03×10^{-4}	8.29×10^{-4}	3.69×10^{-4}	1.12	1.05	1.26	1.09	1
	GIOP	266	0.602	0.947	1.27×10^{-4}	7.73×10^{-5}	7.41×10^{-4}	4.20×10^{-4}	1.02	1.00	1.24	1.12	8
	Huot	266	0.448	1.28	-9.39×10^{-4}	-1.38×10^{-4}	1.08×10^{-3}	$2.64E-4$	0.748	0.786	1.41	1.31	1

Note. Bold text indicates best performance for each skill metric. No. wins (last column) indicates number of statistical tests in which respective dataset outperformed others. *Computed in \log_{10} - \log_{10} space. Difference metrics with correction factor applied.

Does it change the matchup result?

Bias (m^{-1})	Bias' (m^{-1})	MAE (m^{-1})	MAE' (m^{-1})
3.90×10^{-4}	2.61×10^{-4}	8.01×10^{-4}	3.96×10^{-4}
1.52×10^{-4}	9.51×10^{-5}	6.75×10^{-4}	3.86×10^{-4}
-7.38×10^{-4}	-5.49×10^{-4}	9.15×10^{-4}	6.59×10^{-4}

	MAE (m)	bias _s (unitless)	bias _{log} (unitless)	MAE _{log} (unitless)	MAE _s (unitless)	No. wins
	3.96×10^{-4}	1.21	1.12	1.33	1.16	0
	3.86×10^{-4}	1.06	1.03	1.27	1.15	10
	6.59×10^{-4}	0.812	0.834	1.37	1.27	0

	Hust	326	0.699	1.40	-7.38×10^{-4}	-5.49×10^{-4}	9.15×10^{-4}	6.59×10^{-4}	0.812	0.834	1.37	1.27	0
<1.25E-3 m	LH	60	0.225	0.764	6.72×10^{-4}	5.19×10^{-4}							
	GIOP	60	0.049	0.614	2.65×10^{-4}	1.7×10^{-4}							
	Hust	60	0.235	1.01	1.51×10^{-4}	1.47×10^{-4}							
>1.25E-3 m	LH	266	0.548	1.24	3.25×10^{-4}	2.03×10^{-4}							
	GIOP	266	0.002	0.947	1.27×10^{-4}	7.73×10^{-5}							
	Hust	266	0.448	1.28	-9.39×10^{-4}	-1.38×10^{-4}							

bias _{log} (unitless)	bias' _{log} (unitless)	MAE _{log} (unitless)	MAE' _{log} (unitless)	v
1.21	1.12	1.33	1.16	
1.06	1.03	1.27	1.15	
0.812	0.834	1.37	1.27	

Note: Bold text indicates best performance for each skill metric. No. wins: number of matchups where the model outperformed others. *Computed in log-log space. Difference metrics: ...

Prime symbol (') indicates corrected metrics

Visualizing matchups (with uncertainties)

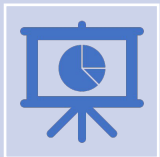


We routinely report validation with one-to-one scatter plots (often in log-transform space)

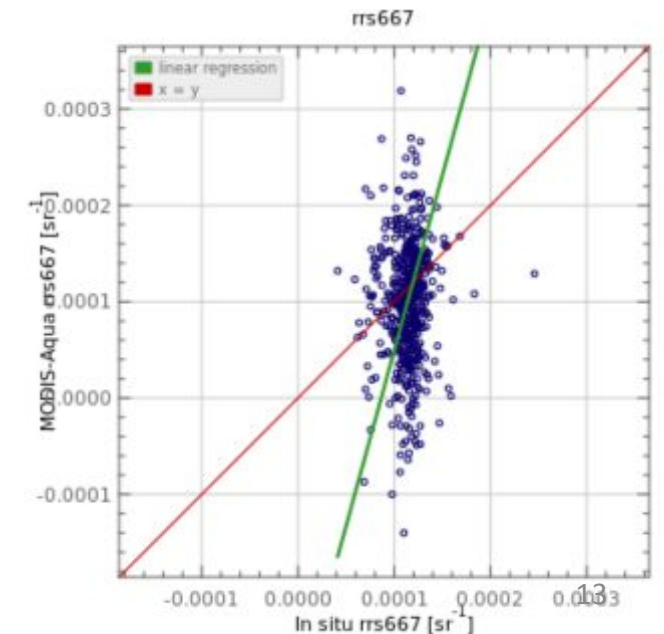
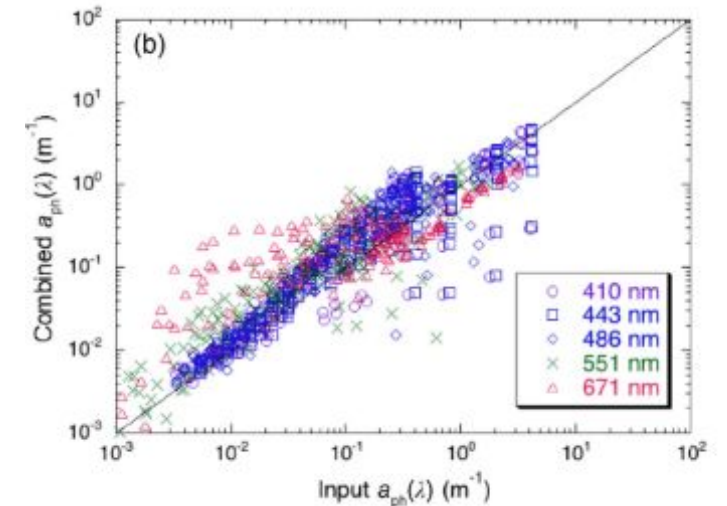


Scatter plots can get messy!

Some folks plot multiple wavelengths or multiple variables



Scatter plots are not that meaningful if the variable has a small dynamic range (e.g. oligotrophic $R_{rs}(670)$)



Visualizing matchups (with uncertainties)

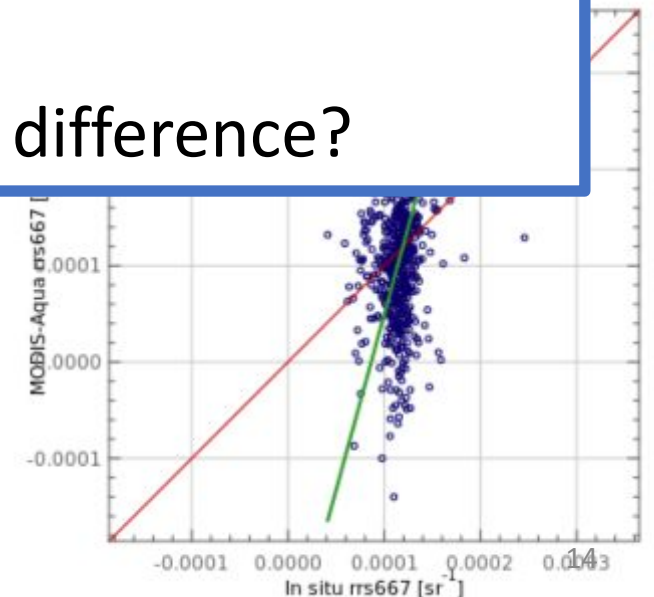
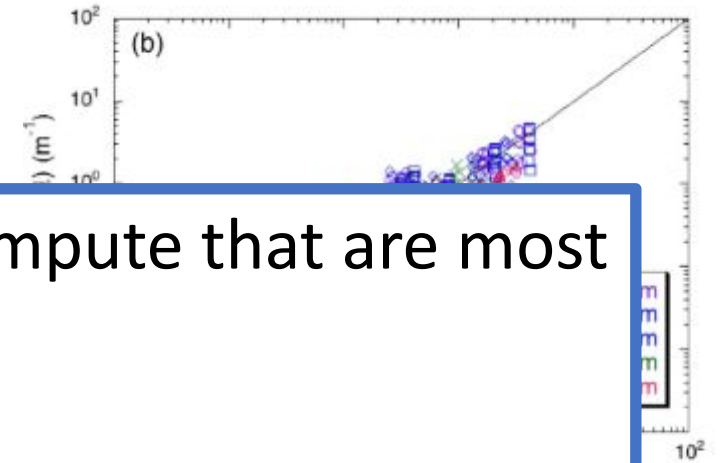
Q. Step back and ask: what is the metric(s) we compute that are most meaningful?

A. Pair-wise difference metrics (e.g. bias, MAE).

So, why don't we create plots that help us visualize difference?



Scatter plots are not that meaningful if the variable has a small dynamic range (e.g. oligotrophic $R_{rs}(670)$)



Zeta score plots

- The zeta-score is the difference between satellite and in situ measurement normalized by total uncertainty

$$\zeta = \frac{M_i - O_i}{\sqrt{u(M_i)^2 + u(O_i)^2}}$$

- Agreement categories can be color-coded

green = *satisfactory* ($|\zeta| < 2$),

yellow = *questionable* ($2 \leq |\zeta| < 3$),

red = *unsatisfactory* ($|\zeta| \geq 3$)

- We can tally the number of data points that fall within each agreement category

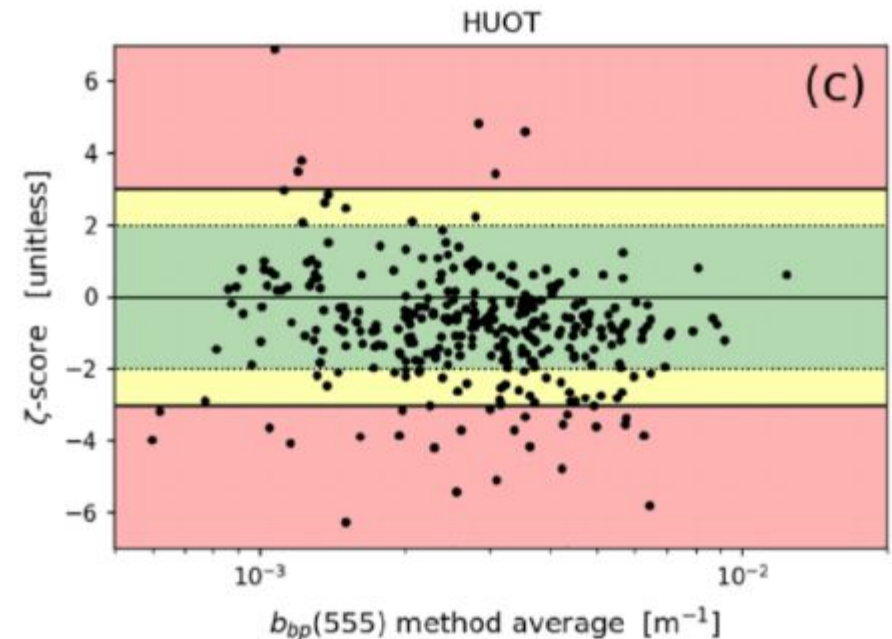
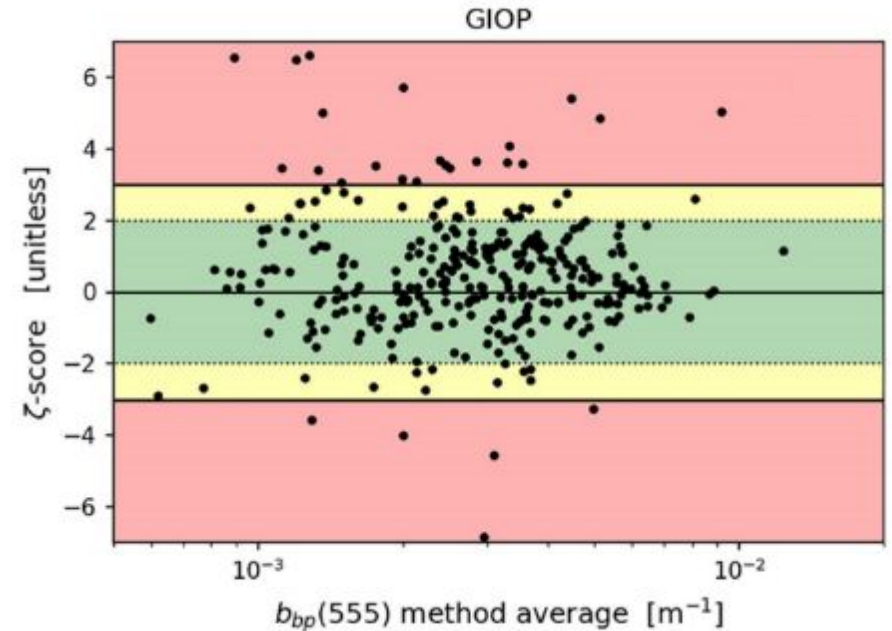


Table 3*Zeta-Score Statistics and Tallies for Three Models: LH, GIOP, and Huot*

$b_{bp}(555)$ range	Model	N	Mean ζ (std)	Mean ζ' (std)	Tally of $ \zeta < 2$	Tally of $ \zeta' < 2$	Tally of $2 \leq \zeta < 3$	Tally of $2 \leq \zeta' < 3$	Tally of $ \zeta \geq 3$	Tally of $ \zeta' \geq 3$	No. wins
All	LH	326	0.888 (1.62)	0.561 (1.36)	249	278	45	22	32	26	0
	GIOP	326	0.348 (1.78)	0.234 (1.55)	265	287	35	14	26	25	8
	Huot	326	-0.814 (1.63)	-0.586 (1.48)	254	272	38	23	34	31	0
$<1.25E-3 \text{ m}^{-1}$	LH	60	2.54 (1.31)	1.92 (1.60)	19	31	19	12	22	17	0
	GIOP	60	1.08 (1.94)	0.739 (1.79)	45	50	8	3	7	7	1
	Huot	60	0.756 (1.22)	0.436 (1.10)	53	55	5	3	2	2	8
$\geq 1.25E-3 \text{ m}^{-1}$	LH	266	0.510 (1.42)	0.252 (1.06)	230	247	26	10	10	9	5
	GIOP	266	0.179 (1.72)	0.119 (1.48)	220	237	28	8	19	18	3
	Huot	266	-1.17 (1.49)	-0.820 (1.45)	201	217	33	20	32	29	0

Note. Bold typeface indicates best performance. Prime symbol (') indicates corrected difference metrics.

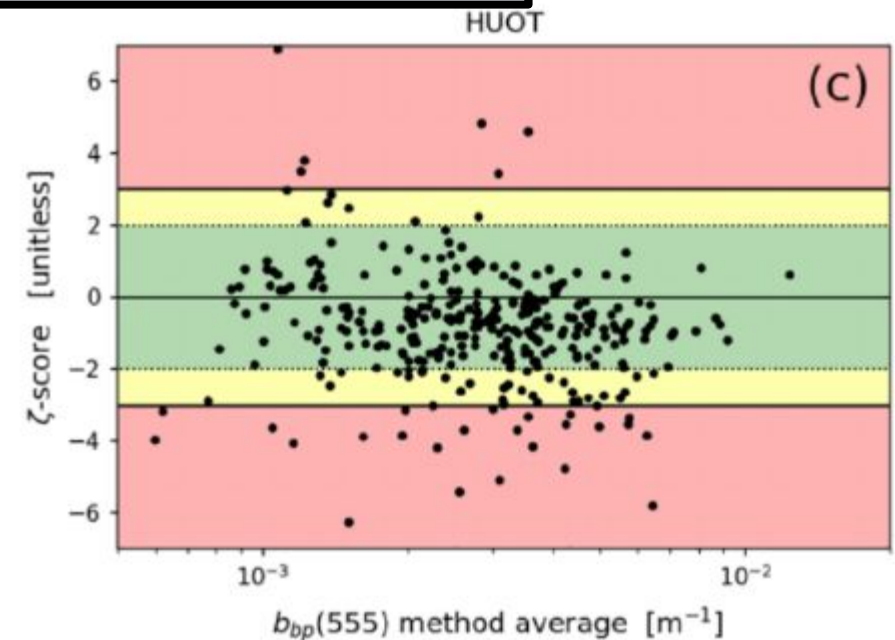
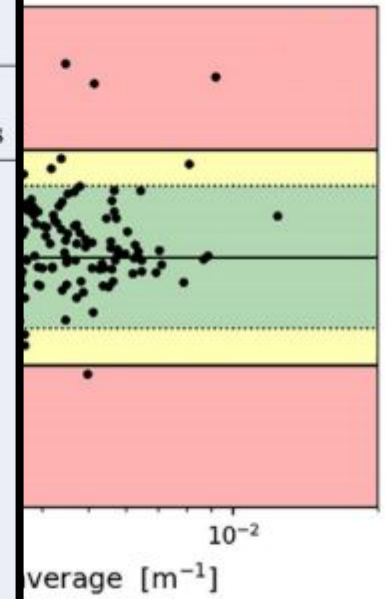
- Agreement categories can be color-coded

green = satisfactory ($|\zeta| < 2$),

yellow = questionable ($2 \leq |\zeta| < 3$),

red = unsatisfactory ($|\zeta| \geq 3$)

- We can tally the number of data points that fall within each agreement category



Summary

1. Uncertainties should be considered during validation – measurements aren't exact
2. Our results show that correcting for uncertainties changes the validation metrics
3. Zeta-score plots show allow us to inspect model residuals
4. Color-coded Zeta-score plots are easy to interpret and may be useful for communicating algorithm performance with end-users.



Some caveats

- If the uncertainties are too large, the corrected validation metrics may have little meaning. *We want to keep uncertainties small*
- We assumed 5% relative uncertainty. Realistic values should be used where possible. Our framework can, however, accommodate alternative input uncertainties if known.
- For retrospective analyses, in situ uncertainty may not be available and we need to make sensible assumptions.

Thanks...

Our results are published here....

AGU ADVANCING
EARTH AND
SPACE SCIENCE

JGR Oceans

RESEARCH ARTICLE
10.1029/2021JC017231

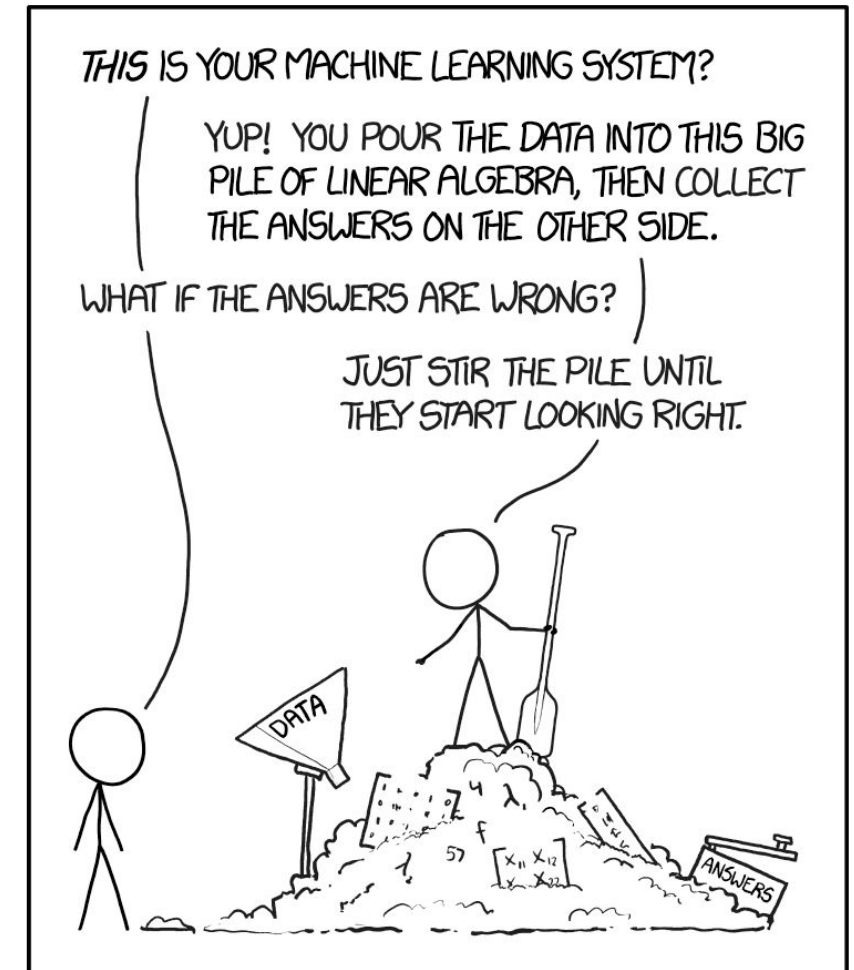

Development and Validation of an Empirical Ocean Color Algorithm with Uncertainties: A Case Study with the Particulate Backscattering Coefficient

Lachlan I. W. McKinna¹, Ivona Cetinić^{2,3}, and P. Jeremy Werdell³

¹Go2Q Pty Ltd, Sunshine Coast, QLD, Australia, ²GESTAR/USRA, Columbia, MD, USA, ³NASA Goddard Flight Center, Greenbelt, MD, USA

Key Points:

- A reflectance line height metric was used as a predictor of the particulate backscattering coefficient at 555 nm
- The degree of overlap metric was used to correct validation skill metrics for measurement



Credit: Academic Twitter

References

- Harmel, R.D., Smith, P.K. and K.W. Migliaccio (2010) Modifying Goodness-of-Fit Indicators to Incorporate Both Measurement and Model Uncertainty in Model Calibration and Validation, *Transactions of the ASABE*, 53(1): 55-63, doi: [10.13031/2013.29502](https://doi.org/10.13031/2013.29502)
- Merchant, C.J. and 17 co-authors (2017) Uncertainty information in climate data records from Earth observation, *Earth Syst. Sci. Data*, 9, 511 – 527, doi: [10.5194/essd-9-511-2017](https://doi.org/10.5194/essd-9-511-2017)
- McKinna, L.I.W., Cetinic, I., and P.J. Werdell (2021) Development and Validation of an Empirical Ocean Color Algorithm with Uncertainties: A Case Study with the Particulate Backscattering Coefficient, *J Geophys Res: Oceans*, 126, doi: [10.1029/2021JC017231](https://doi.org/10.1029/2021JC017231)
- McKinna, L.I.W., Cetinic, I., Chase, A.P. and P.J. Werdell (2019) Approach for Propagating Radiometric Data Uncertainties Through NASA Ocean Color Algorithms, *Front. Earth Sci.*, 7:176, doi: [10.3389/feart.2019.00176](https://doi.org/10.3389/feart.2019.00176)
- Seegers, B.N., Stumpf, R.P., Schaffer B.A., Loftin, K.A., and P.J. Werdell (2018) Performance metrics for the assessment of satellite data products: An ocean color case study, *Opt. Express*, 26(6), 7404-7422, doi: [10.1364/OE.26.007404](https://doi.org/10.1364/OE.26.007404)

Extra material

Final thought from Merchant et al (2017)

It is clear that developing and validating uncertainty estimates involves effort comparable to developing the retrieval itself.

Earth Syst. Sci. Data, 9, 511–527, 2017
<https://doi.org/10.5194/essd-9-511-2017>
© Author(s) 2017. This work is distributed under the Creative Commons Attribution 3.0 License.



Earth System
Science
Data
Open Access

Uncertainty information in climate data records from Earth observation

Christopher J. Merchant^{1,2}, Frank Paul³, Thomas Popp⁴, Michael Ablain⁵, Sophie Bontemps⁶, Pierre Defourny⁶, Rainer Hollmann⁷, Thomas Lavergne⁸, Alexandra Laeng⁹, Gerrit de Leeuw¹⁰, Jonathan Mittaz^{1,11}, Caroline Poulsen¹², Adam C. Povey¹³, Max Reuter¹⁴, Shubha Sathyendranath¹⁵, Stein Sandven¹⁶, Viktoria F. Sofieva¹⁰, and Wolfgang Wagner¹⁷

Model evaluation matrix – what can we achieve?

Table 4. Model evaluation matrix for appropriate model performance conclusions in model calibration/validation considering both measurement uncertainty and prediction uncertainty.

Case	Uncertainty in Measured Data	Uncertainty in Model Predictions	Overall Model Performance Conclusions Based on Model Accuracy (goodness-of-fit) and Model Precision	
			“Good” Indicator Values	“Unsatisfactory” Indicator Values
			1	High
2	High	High	Low model precision, but high measurement uncertainty prevents definitive model accuracy conclusion in spite of good fit indication.	Unsatisfactory model performance due to low precision and poor accuracy.
3	Low	High	Low model precision, but good model accuracy.	Unsatisfactory model performance due to low precision and poor accuracy.
4	Low	Low	Good model performance in terms of high precision and good accuracy.	Unsatisfactory model performance due to poor accuracy in spite of high model precision.

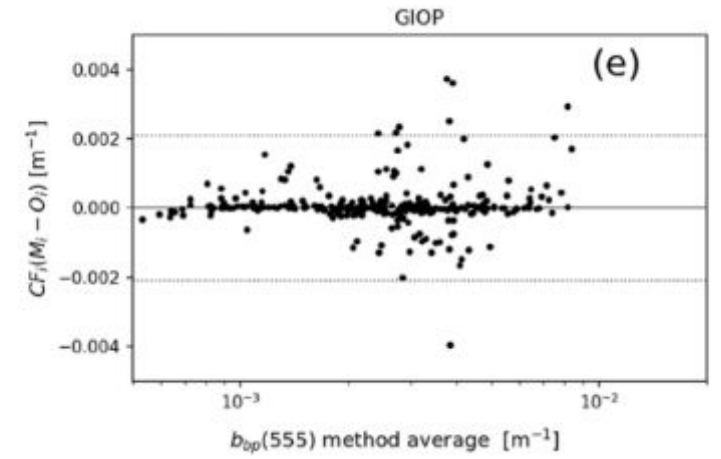
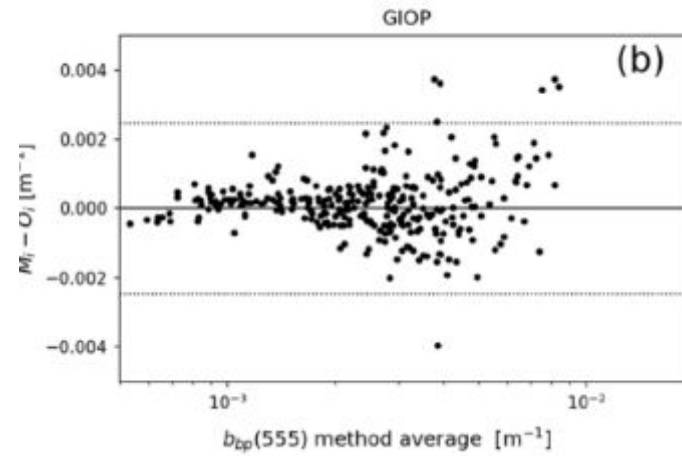
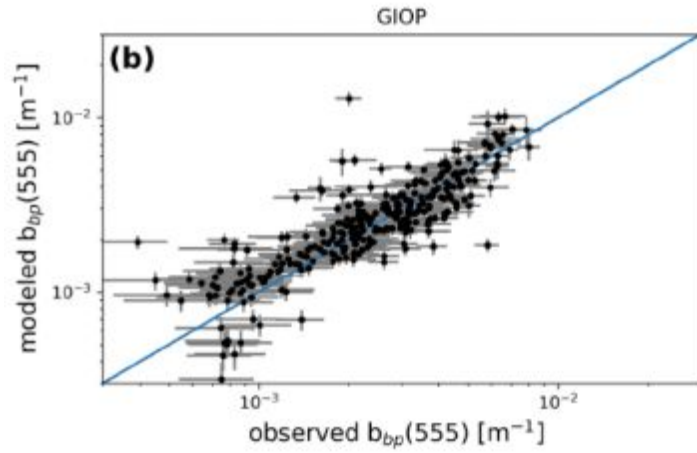
Bland-Altman Plots (residual plots)

\log_{10} - \log_{10} scatter

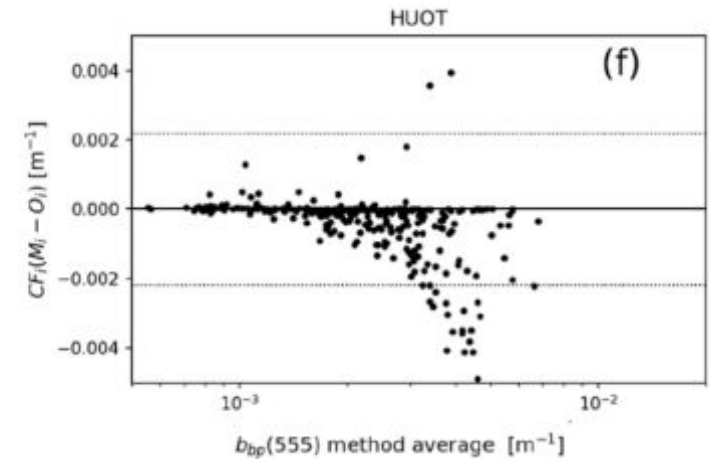
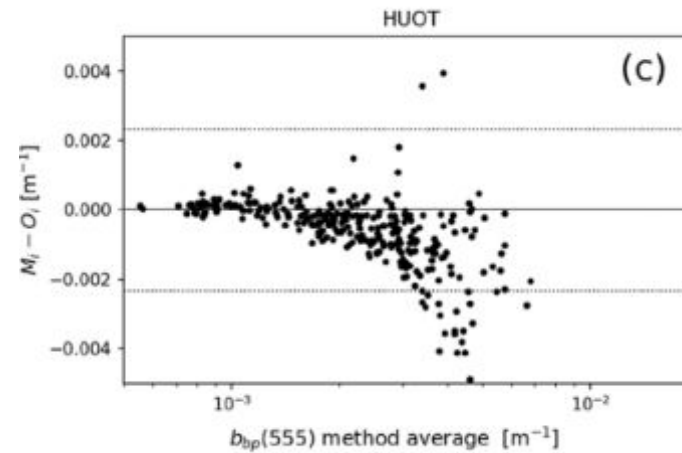
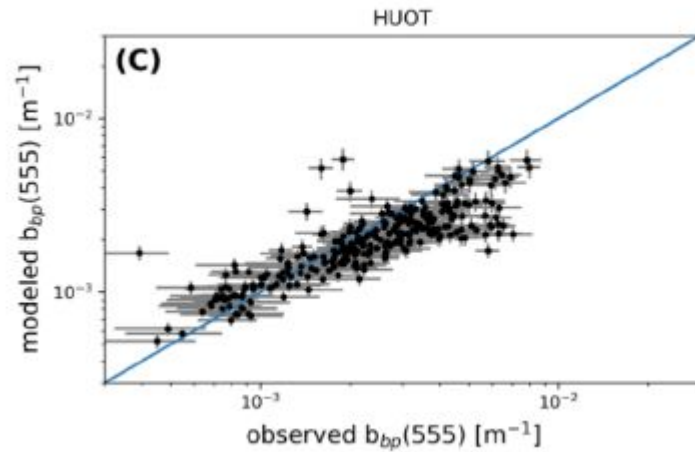
Bland-Altman

Bland-Altman
(with corrected differences)

Model 1

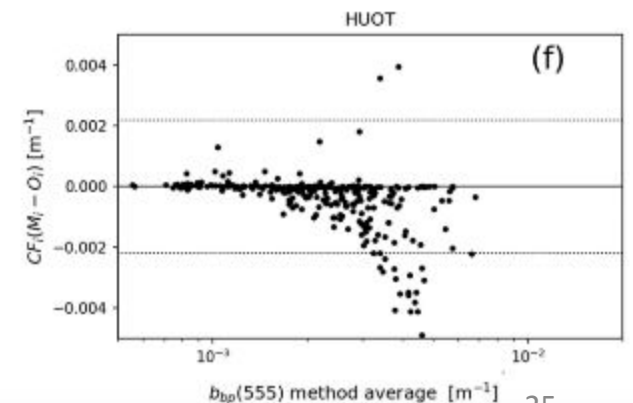
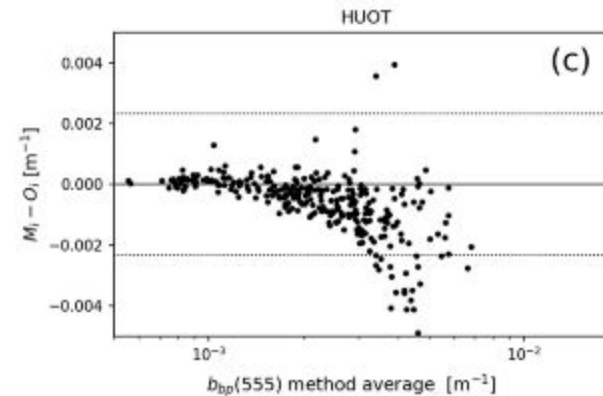
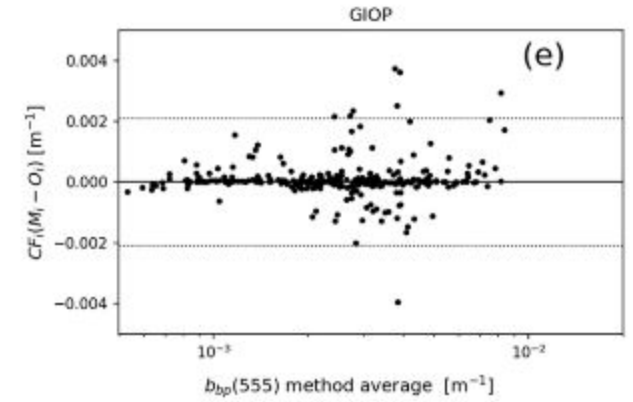
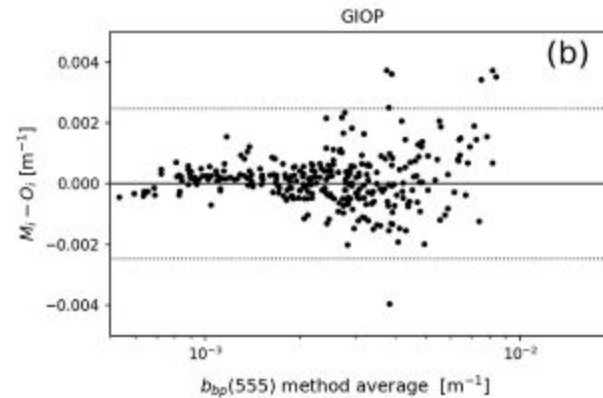
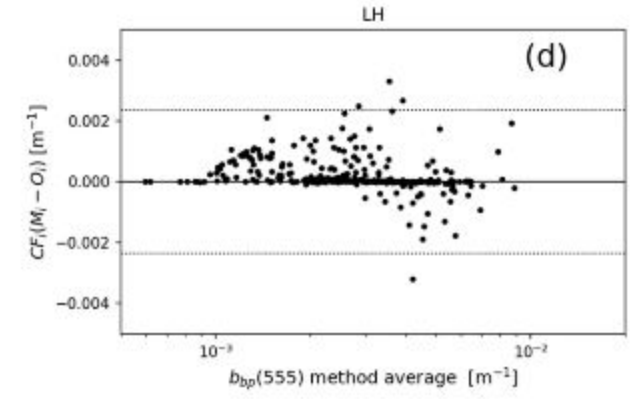
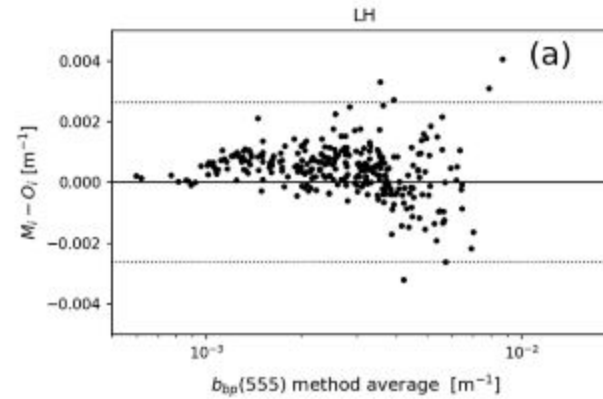


Model 2



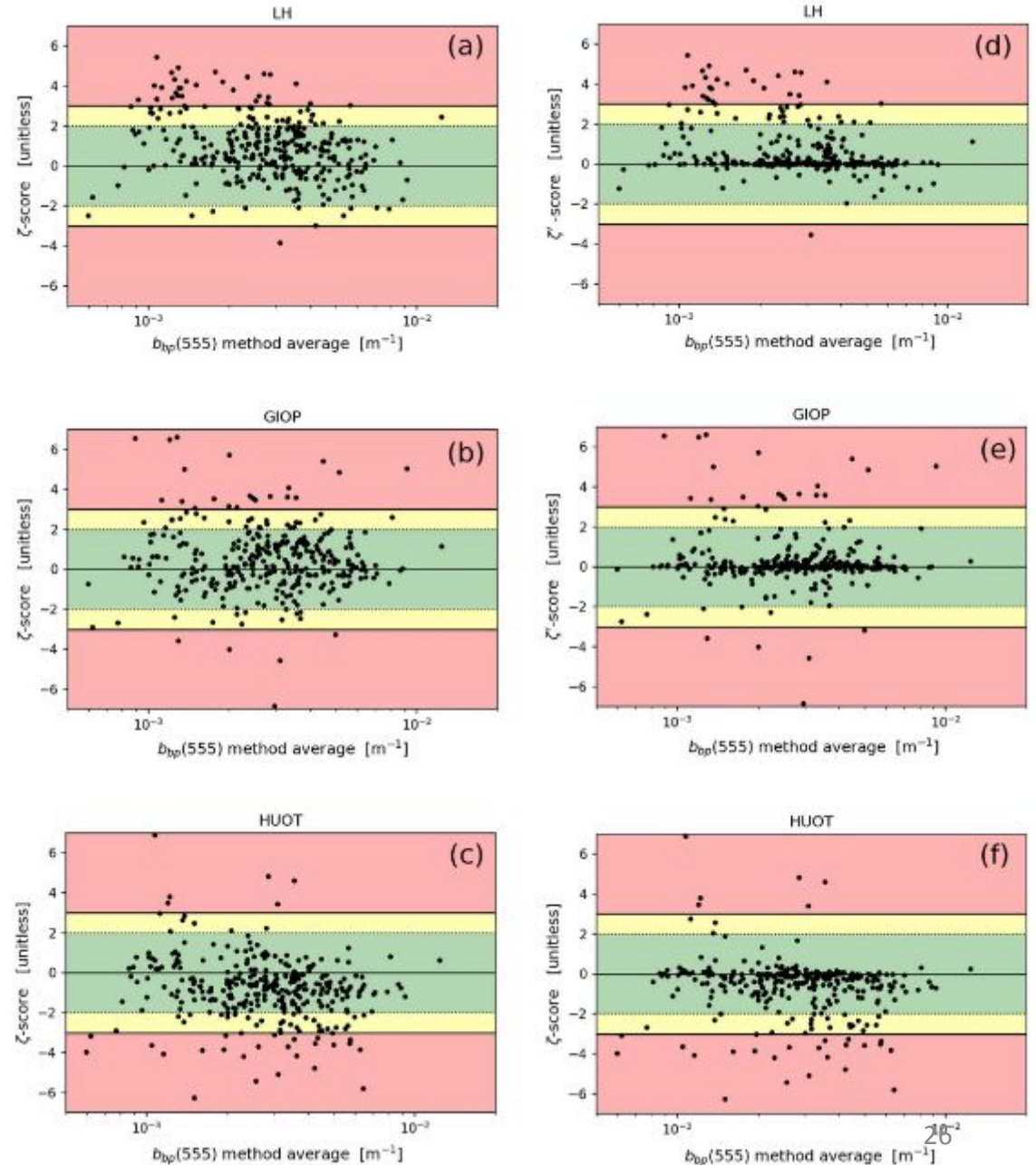
Bland-Altman Plots

- Difference on the y-axis
- Method average on x-axis



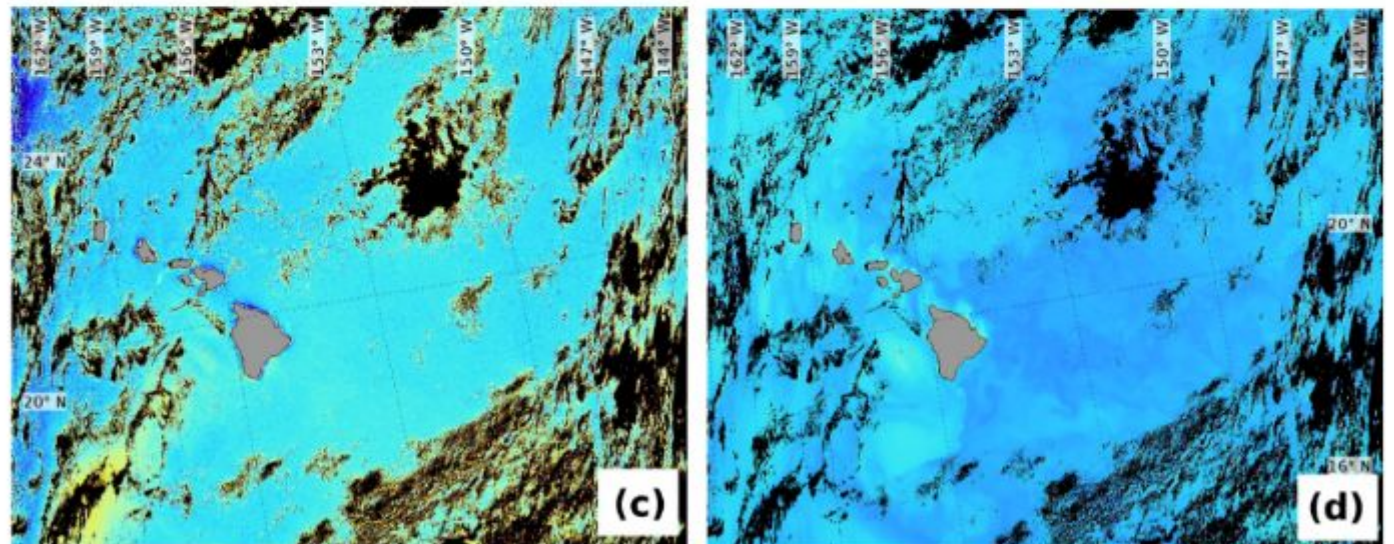
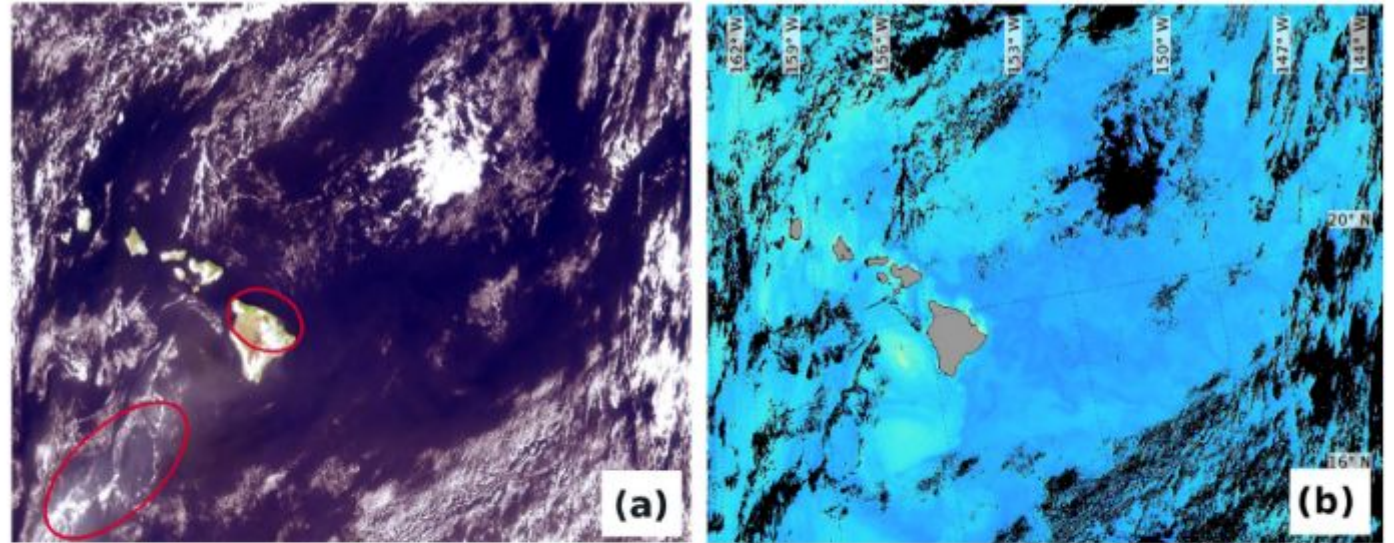
Zeta-score plots

- Zeta-score on y-axis
- Method average on x-axis

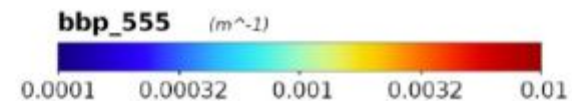


SeaWiFS plots

Derived b_{bp} (555) in waters adjacent to Hawaii (1st Dec 2000)

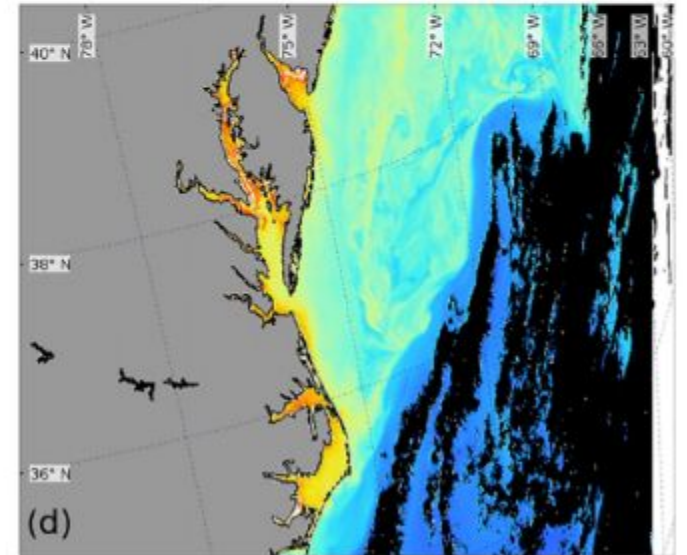
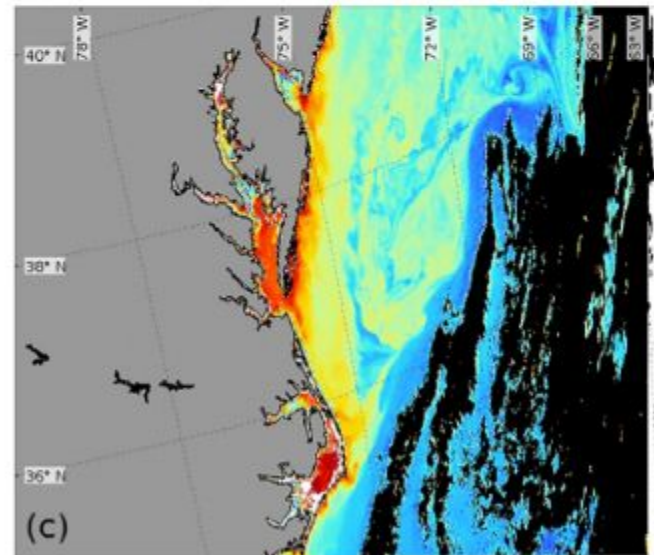
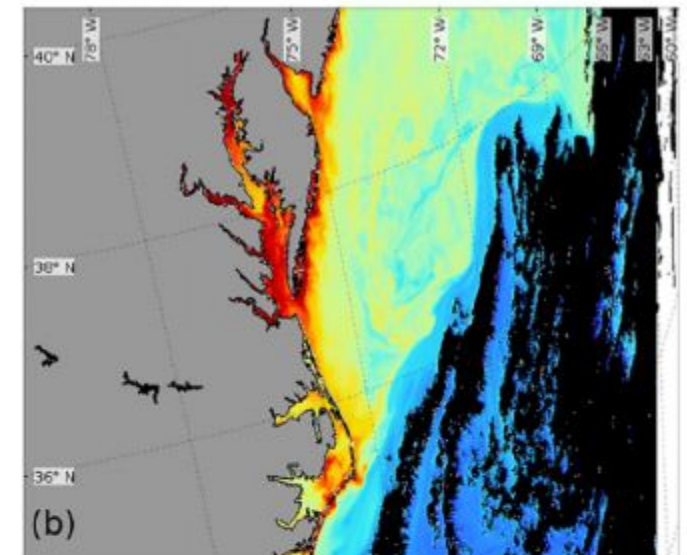
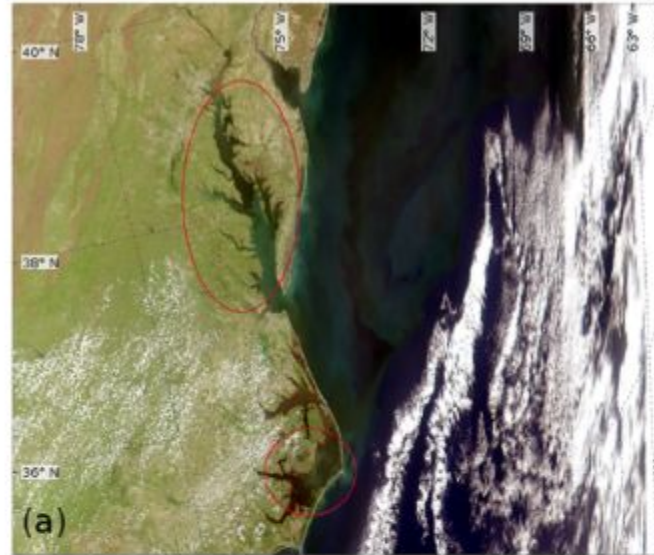


- (a) RGB
- (b) LH empirical model
- (c) GIOP
- (d) Huot empirical model

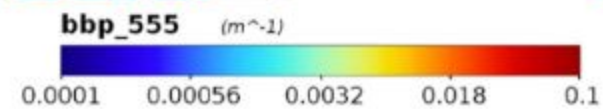


SeaWiFS plots

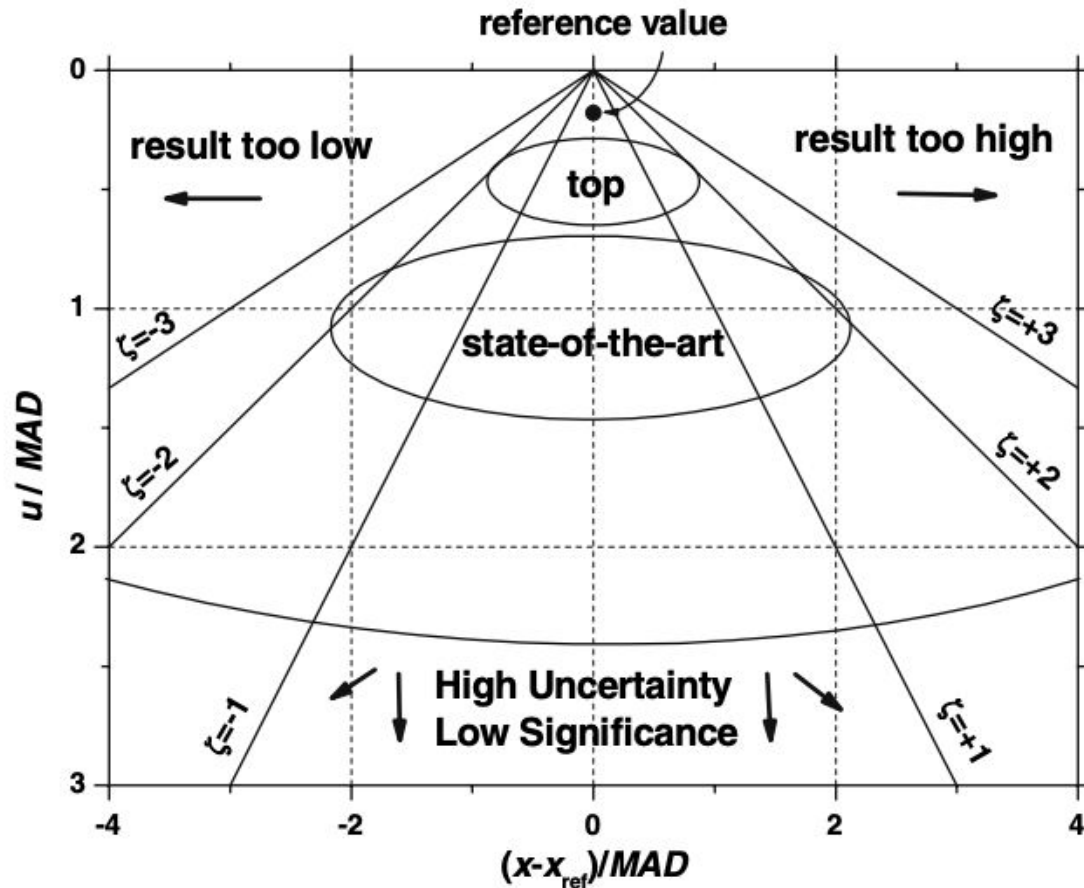
Derived b_{bp} (555) in
Mid-Atlantic Bight
(28th Apr 2003)



- (a) RGB
- (b) LH empirical model
- (c) GIOP
- (d) Huot empirical model



PomPlots overview



- Pair-wise comparisons closest to the apex (0,0) are best
- Bias (left-to-right) can be interpreted
- Relative uncertainty (vertical scale) helps us interpret how useful a data point is
- ζ -score contours are also shown

PomPlot of $b_{bp}(443)$ reveals a lot!

