CERES-only Machine Learning Data Product: Overview and Recent Progress

Takmeng Wong[1], Bijoy Vengasseril Thampi[2], and the CERES Data Management Team

[1]NASA Langley Research Center, Hampton, Virginia

[2]AMA, Hampton, Virginia

40th CERES Science Team Meeting
Hampton, Virginia
May 14-16, 2024

Outline

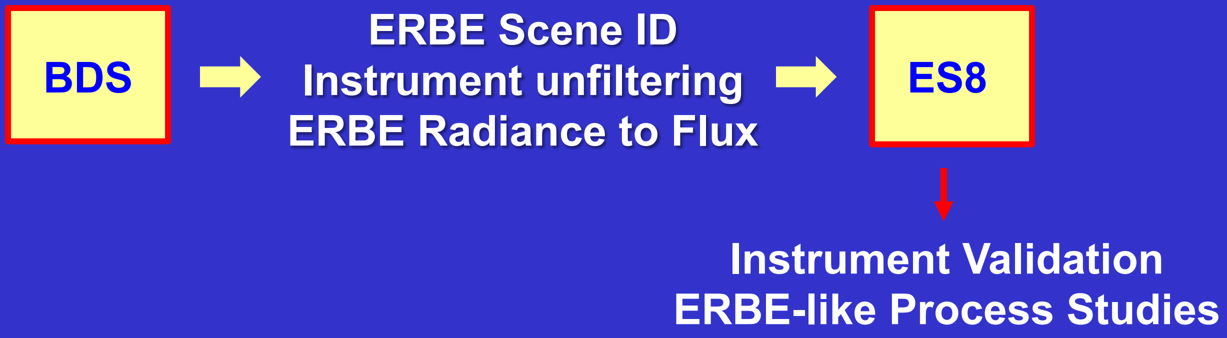
- What is the CERES-only Machine Learning (ML) Data Product?
- Differences between Edition4 ERBE-like and Edition5 ML
- Current Status and Plans
- ML Scene ID Algorithm and Results
- ML Radiance to Flux Algorithm and Early Results
- Summary

What is the CERES-only ML Data Product?

- The CERES-only Machine Learning (ML) data product is a level 2 data product, generated without any imager (i.e., MODIS, VIIRS, …) data.
- It will replace the CERES ERBE-like data product, which itself is also a CERES-only imager-less data set, in the next CERES Edition5 data release.
- The goal of the CERES-only ML data product is to improve the quality of the CERES-only data by replacing the 35 years old legacy ERBE algorithms with modern machine learning techniques.
- There will be two data sets in the CERES-only ML level 2 data product.
  - ➢ CERES-only ML Radiances (MLR) data (radiance only)
  - ➢ CERES-only ML Fluxes (MLF) data (both radiance and fluxes)
- The MLR data will be used internally to support CERES instrument calibration activities.
- The MLR data will be used as CERES TISA data gap filler when CERES SSF TOA fluxes are unavailable due to lack of imager data.

CERES-only Data: Edition4 vs Edition5

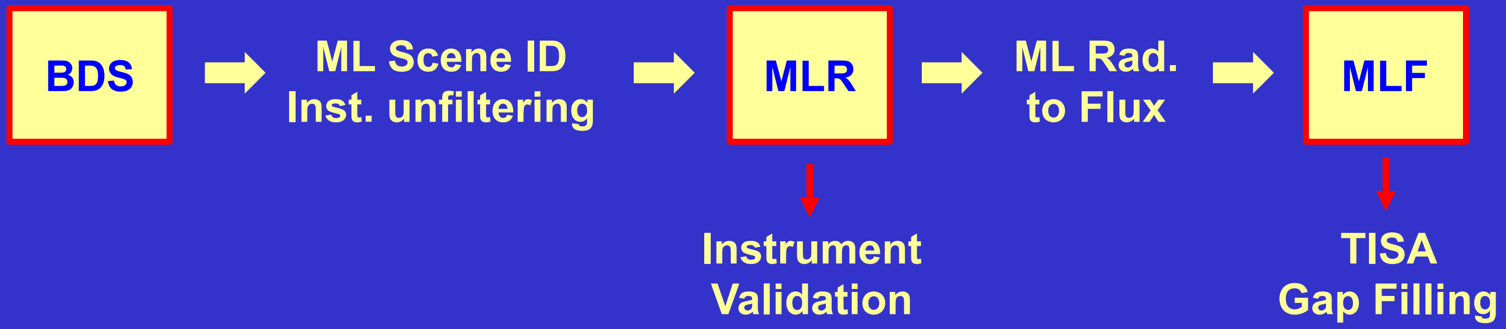
Edition4: Legacy ERBE Algorithm



BDS → ERBE Scene ID
Instrument unfiltering
ERBE Radiance to Flux → ES8

ES8 → Instrument Validation
ERBE-like Process Studies

Edition5: New Machine Learning Algorithm



BDS → ML Scene ID
Inst. unfiltering → MLR → ML Rad.
to Flux → MLF

MLR → Instrument
Validation

MLF → TISA
Gap Filling

Current Status of the CERES-only ML Data Product

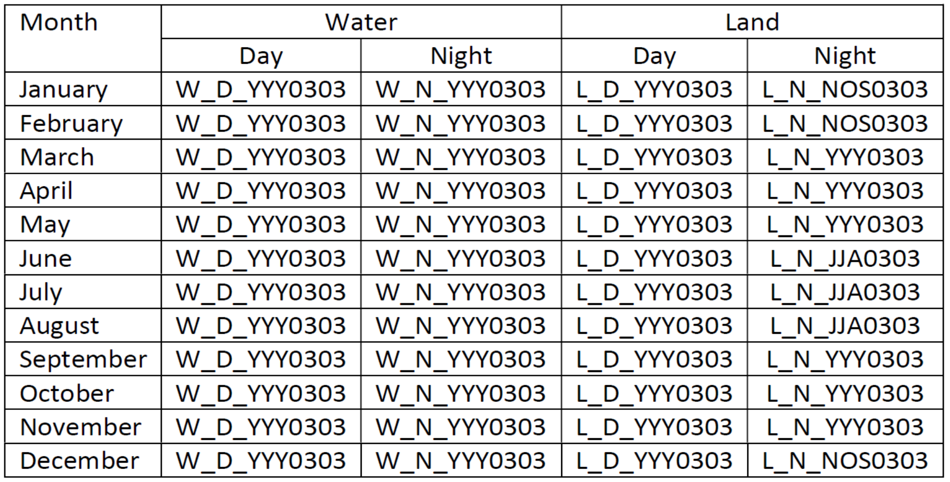
- MLR (radiances) data set:
  - ➢ Science software, data set format and variable listing completed in 2/24
  - ➢ CERES Data Management Team (DMT) is currently converting the Science software into production code (estimated completion by Fall 2024)
  - ➢ Final production code will be tested and certified to operate in the ASDC environment before data production begin (TBD)

- MLF (fluxes) data set.
  - ➢ Science algorithm started in 3/24 (estimated completion by 3/25)
  - ➢ Data set format and variable listing (estimated completion by 3/25)
  - ➢ Conversion to production code by CERES DMT (estimated completion by end of Fall 2025)
  - ➢ Final production code at ASDC (TBD)

CERES ML Scene ID Algorithm

- CERES ML Scene ID Algorithm uses a Machine Learning decision tree classification technique known as the Random Forrest (RF) Method
- RF method classifies each CERES footprint as clear or cloudy (a binary operation)
- RF model is trained (i.e., deep learning) using Ed4 SSF data by matching the clear (99.9% clear) and cloudy footprints with the following variables:
  - Daytime: SW and LW radiance, solar zenith, viewing zenith, relative azimuth
  - Night-time: LW radiance, viewing zenith, and latitude
- Separate RF models: water, land, desert, and ice/snow; day and night
- There are three possible outcomes for the RF determined clear-sky data:
  - Exact Match (EM) when RF classified SSF clear footprint as clear footprint
  - False Positive (FP) when RF classified SSF cloudy footprint as clear footprint
  - False Negative (FN) when RF classified SSF clear footprint as cloudy footprint
- Trial and error method is used to find the best model with the least amount of FP clear footprints; these FP footprints are makeup by low COD and/or low CF footprints that are difficult to identify correctly in RF model.
- The best RF model is tested using data from 2015 calendar year

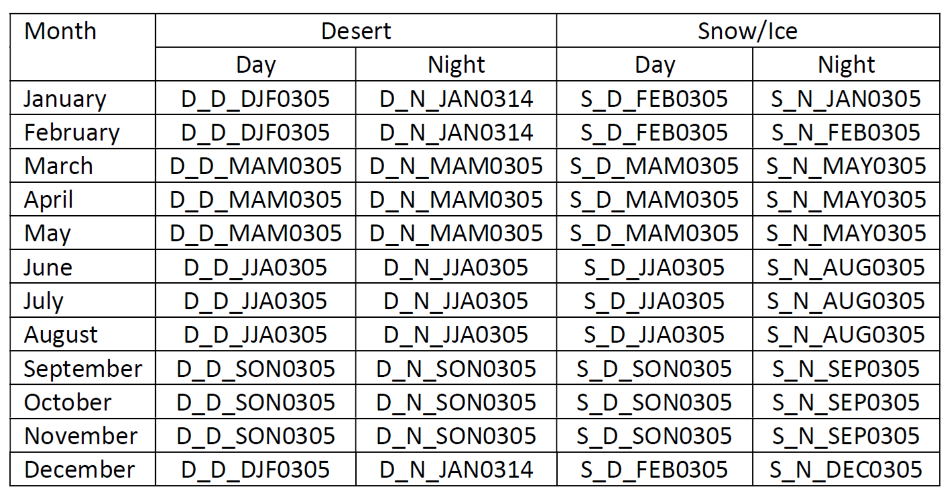
CERES ML Scene ID RF Models



Month	Water		Land	
	Day	Night	Day	Night
January	W_D_YYY0303	W_N_YYY0303	L_D_YYY0303	L_N_NOS0303
February	W_D_YYY0303	W_N_YYY0303	L_D_YYY0303	L_N_NOS0303
March	W_D_YYY0303	W_N_YYY0303	L_D_YYY0303	L_N_YYY0303
April	W_D_YYY0303	W_N_YYY0303	L_D_YYY0303	L_N_YYY0303
May	W_D_YYY0303	W_N_YYY0303	L_D_YYY0303	L_N_YYY0303
June	W_D_YYY0303	W_N_YYY0303	L_D_YYY0303	L_N_JJA0303
July	W_D_YYY0303	W_N_YYY0303	L_D_YYY0303	L_N_JJA0303
August	W_D_YYY0303	W_N_YYY0303	L_D_YYY0303	L_N_JJA0303
September	W_D_YYY0303	W_N_YYY0303	L_D_YYY0303	L_N_YYY0303
October	W_D_YYY0303	W_N_YYY0303	L_D_YYY0303	L_N_YYY0303
November	W_D_YYY0303	W_N_YYY0303	L_D_YYY0303	L_N_YYY0303
December	W_D_YYY0303	W_N_YYY0303	L_D_YYY0303	L_N_NOS0303

Water and Land Model (6 Models)

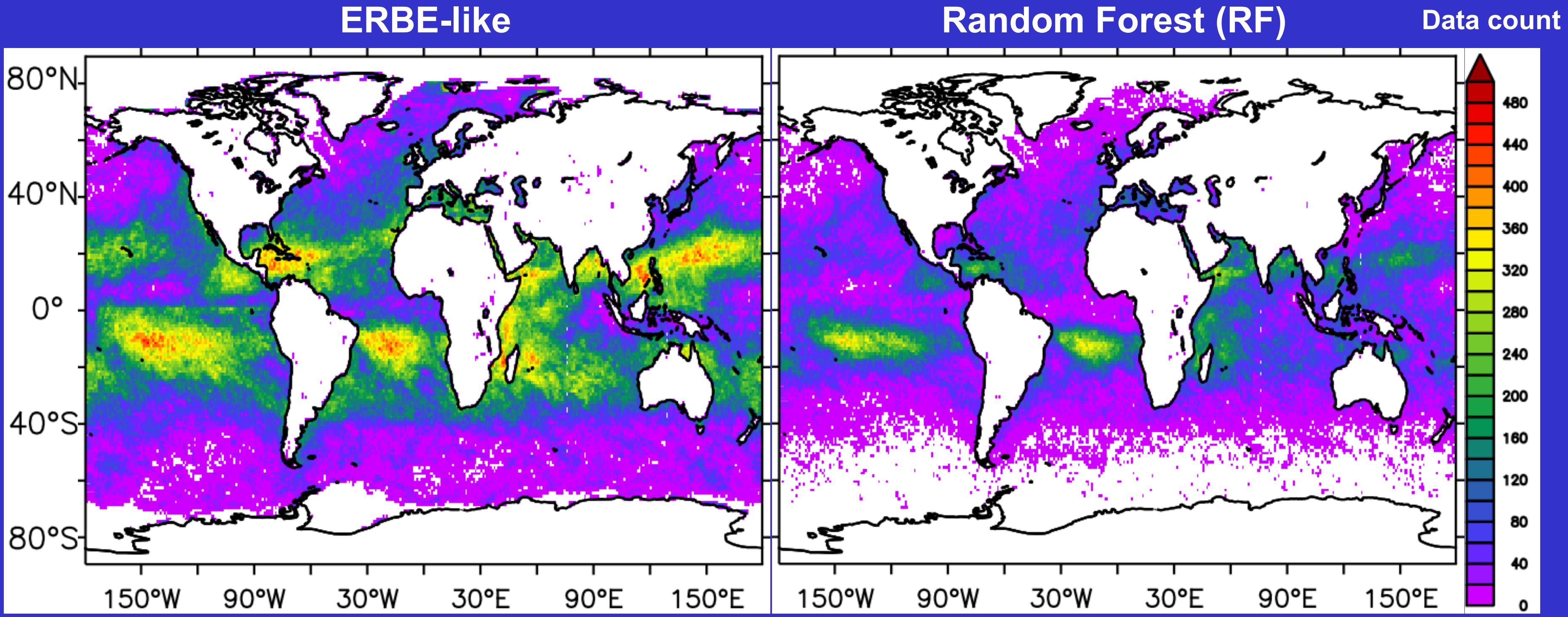
Desert and Snow/ice Model (18 Models)



Month	Desert		Snow/Ice	
	Day	Night	Day	Night
January	D_D_DJF0305	D_N_JAN0314	S_D_FEB0305	S_N_JAN0305
February	D_D_DJF0305	D_N_JAN0314	S_D_FEB0305	S_N_FEB0305
March	D_D_MAM0305	D_N_MAM0305	S_D_MAM0305	S_N_MAY0305
April	D_D_MAM0305	D_N_MAM0305	S_D_MAM0305	S_N_MAY0305
May	D_D_MAM0305	D_N_MAM0305	S_D_MAM0305	S_N_MAY0305
June	D_D_JJA0305	D_N_JJA0305	S_D_JJA0305	S_N_AUG0305
July	D_D_JJA0305	D_N_JJA0305	S_D_JJA0305	S_N_AUG0305
August	D_D_JJA0305	D_N_JJA0305	S_D_JJA0305	S_N_AUG0305
September	D_D_SON0305	D_N_SON0305	S_D_SON0305	S_N_SEP0305
October	D_D_SON0305	D_N_SON0305	S_D_SON0305	S_N_SEP0305
November	D_D_SON0305	D_N_SON0305	S_D_SON0305	S_N_SEP0305
December	D_D_DJF0305	D_N_JAN0314	S_D_FEB0305	S_N_DEC0305

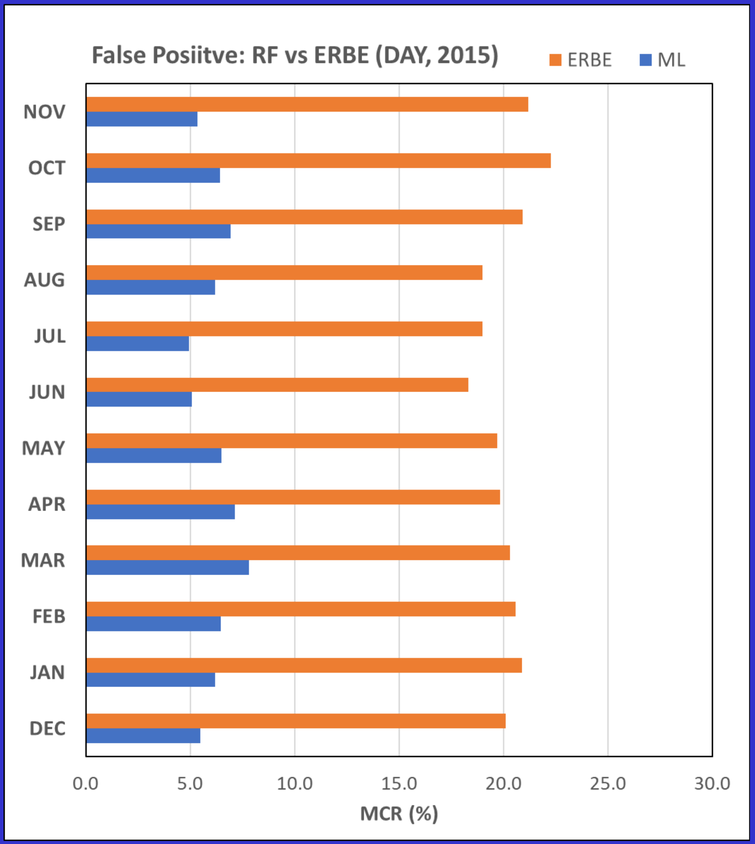
CERES ML RF Water FP Results

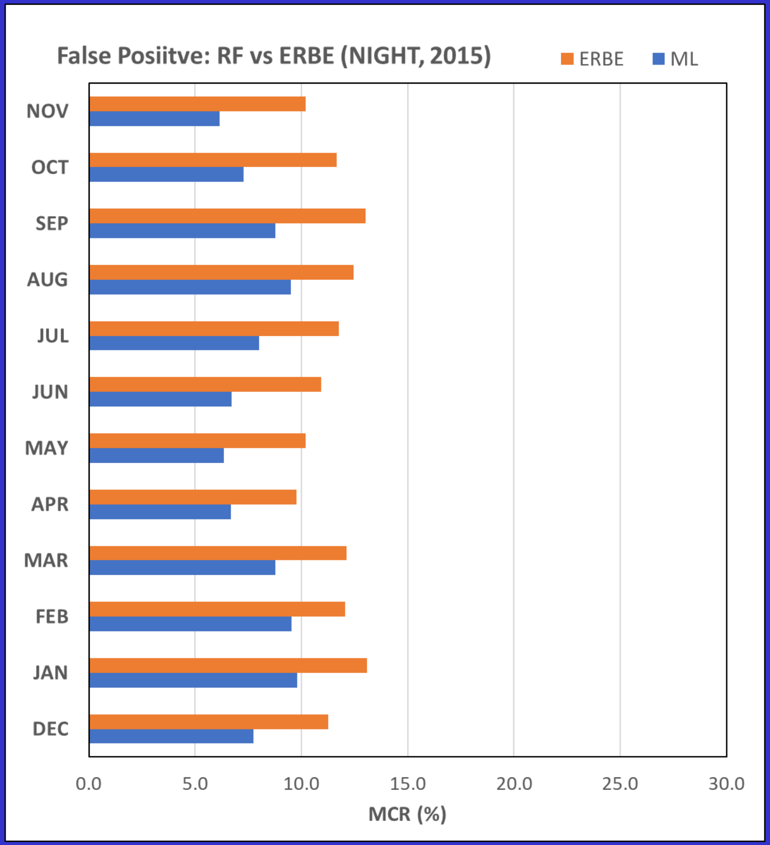
ERBE-like vs RF, Water, Daytime, 04/2015, FP Clear Footprint Count



SSF Clear Footprints	ERBE-like Clear Footprints	ERBE-like False Positive (FP) Clear Footprints	RF Clear-sky Footprints	RF False Positive Clear Footprints
657,066	4,470,146	3,877,062	1,817,448	1,396,354

CERES ML RF Water FP Result (Cont.)



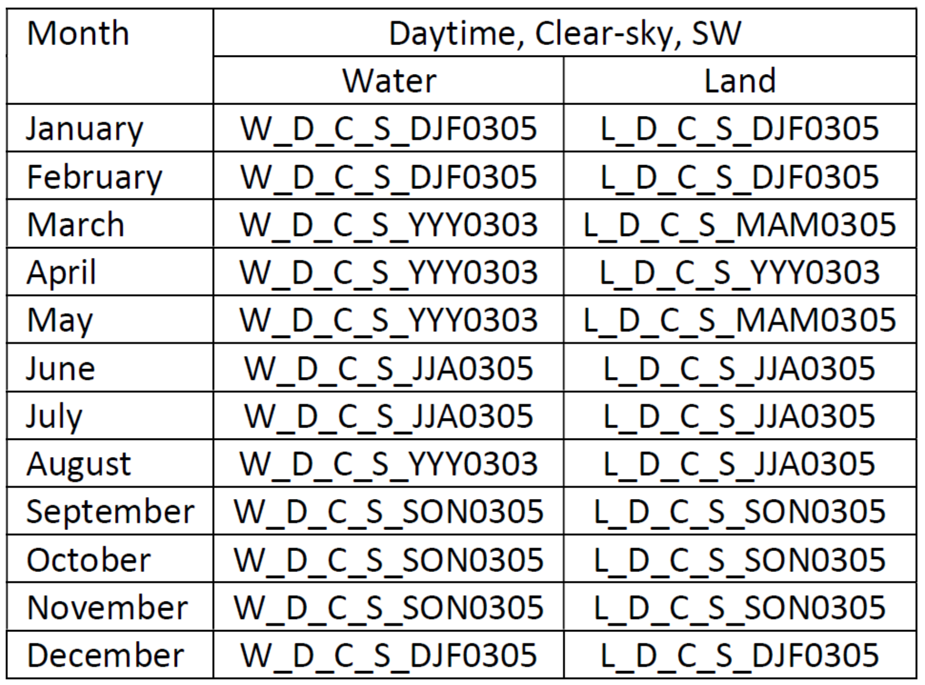


- ML Scene ID RF algorithm is superior to the Legacy ERBE Scene ID algorithm for both day and night.
- Daytime RF models perform better than night-time RF models.

CERES ML Radiance to Flux Algorithm

- CERES ML Radiance to Flux Algorithm uses a Machine Learning technique known as Artificial Neural Network (ANN) method
- ANN method converts the satellite altitude radiance in each CERES footprint into TOA flux
- ANN clear-sky and cloudy-sky models are trained (i.e., deep learning) using Ed4 SSF clear (99.9% clear) and cloudy footprints, respectively, with the following variables:
  - ➢ SW Flux: SW and LW radiance, solar zenith, viewing zenith, relative azimuth
  - ➢ LW/WN Flux: LW/WN radiance, viewing zenith, precipitable water, latitude…
- Separate ANN models: water, land, desert, and ice/snow; day and night, clear and cloudy
- Trail and error method is used to tune the ANN models in order to minimize both the global mean bias and global mean RMS differences between ANN (model) and SSF (truth) data.
- The best ANN model is tested using data from 2015 calendar year

CERES ML Clear-sky Radiance to Flux ANN Models



Month	Daytime, Clear-sky, SW	
	Water	Land
January	W_D_C_S_DJF0305	L_D_C_S_DJF0305
February	W_D_C_S_DJF0305	L_D_C_S_DJF0305
March	W_D_C_S_YYY0303	L_D_C_S_MAM0305
April	W_D_C_S_YYY0303	L_D_C_S_YYY0303
May	W_D_C_S_YYY0303	L_D_C_S_MAM0305
June	W_D_C_S_JJA0305	L_D_C_S_JJA0305
July	W_D_C_S_JJA0305	L_D_C_S_JJA0305
August	W_D_C_S_YYY0303	L_D_C_S_JJA0305
September	W_D_C_S_SON0305	L_D_C_S_SON0305
October	W_D_C_S_SON0305	L_D_C_S_SON0305
November	W_D_C_S_SON0305	L_D_C_S_SON0305
December	W_D_C_S_DJF0305	L_D_C_S_DJF0305

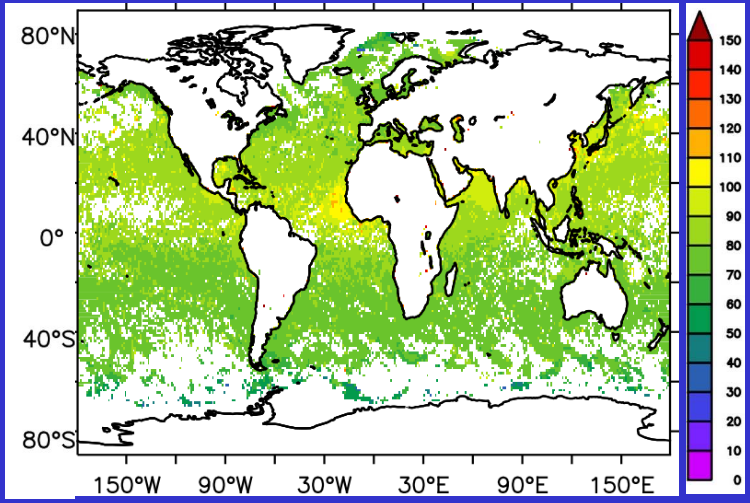
← **Daytime, Clear-sky, SW, Water and Land Model (9 Models)**

- Other ANN clear-sky models for daytime SW (desert and snow/ice), daytime and night-time LW (water, land, desert, snow/ice), daytime and night-time WN (water, land, desert, snow/ice) are currently at work
- Works for ANN cloudy-sky models will begin after completion of all clear-sky models

CERES ML Clear-sky Water ANN Results

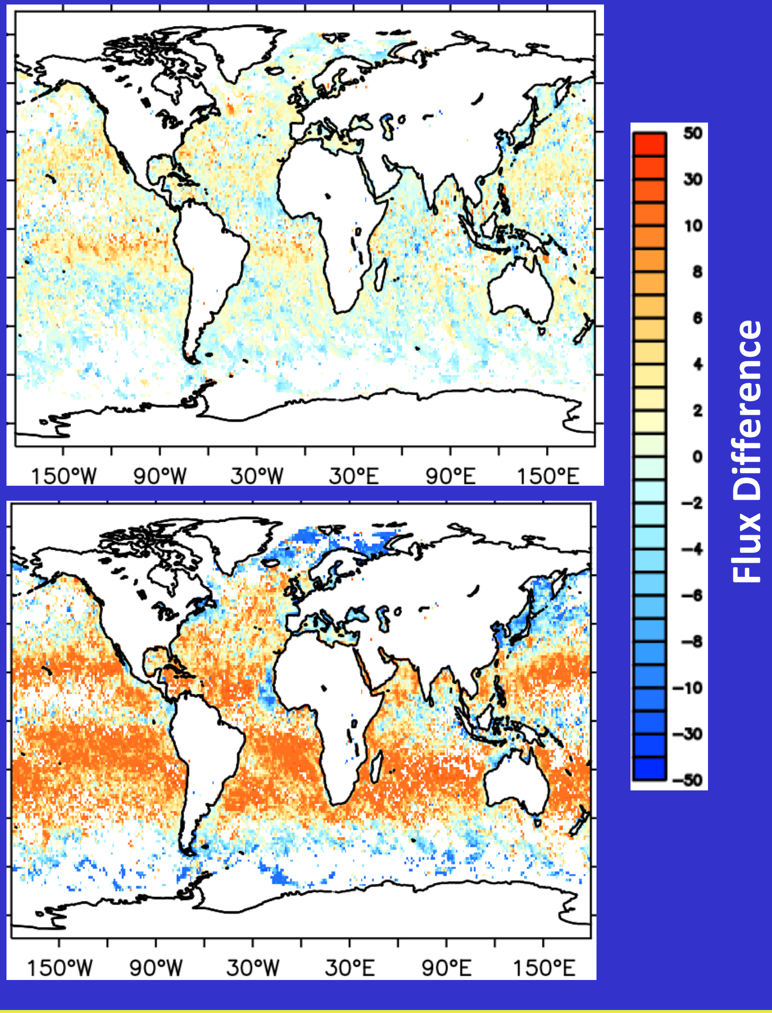
ERBE-like vs ANN, Water, Daytime, Clear-sky, SW Flux, 04/2015

SSF Clear-sky SW Flux



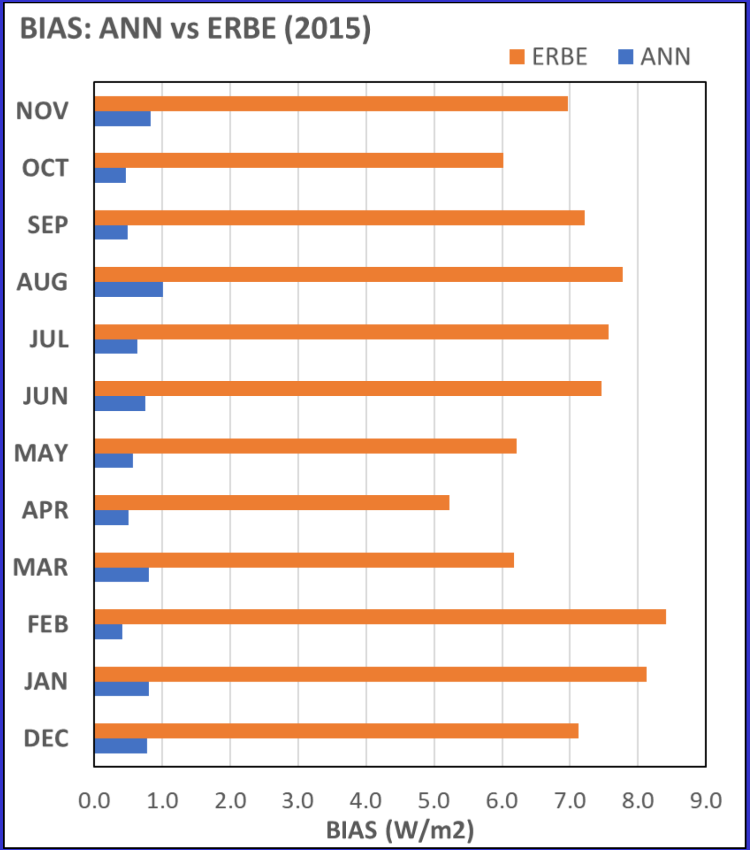
ANN minus SSF

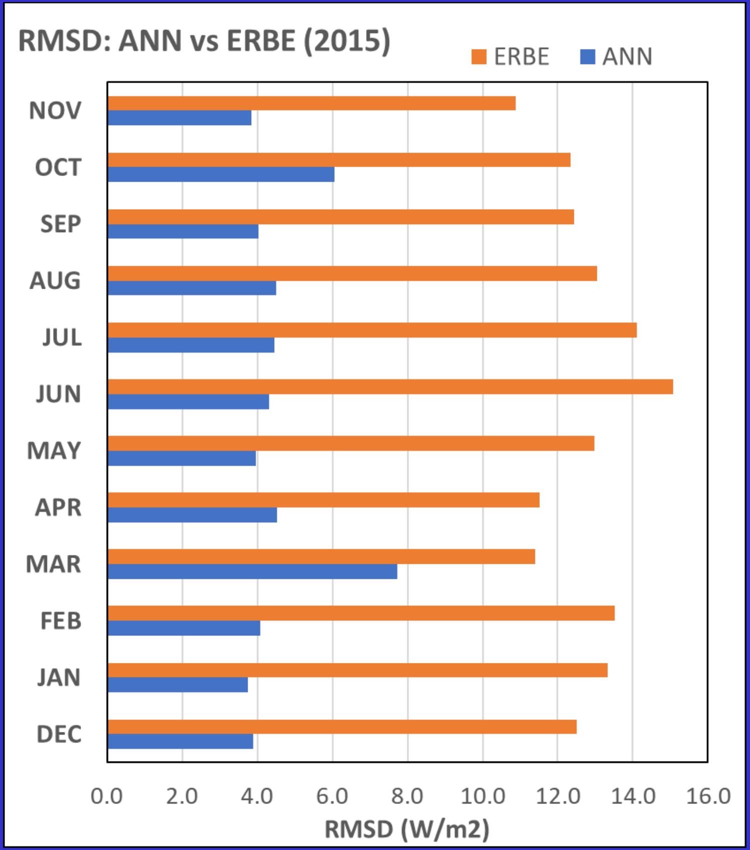
ERBE-like minus SSF



CERES ML Clear-sky Water ANN Results (Cont.)

ERBE-like vs ANN, Ocean, Daytime, Clear-sky, SW Fluxes, 2015





- ANN method is superior to the ERBE method and produces better bias and RMS difference for every calendar month.

Summary

- The new CERES-only ML algorithms (both scene ID and radiance to flux) are shown to be superior to the legacy ERBE algorithms.
- The CERES MLR science algorithm is completed and is currently being converted to production software by the CERES DMT.
- The CERES MLF science algorithm is currently being work on and is expected to be completed early next year.
- This new data product should improve the quality of the CERES-only data products.
- The MLR data will be use internally for CERES instrument validation.
- The MLF data will be useful for TISA data gap filling when SSF TOA fluxes are unavailable due to lack of imager data.