Ingesting Data to Ecosystem and Machine-Learning Models

Yun Li[1]

Tianyu Zhou[1], Zhangxian Ouyang[1], Wei-Jun Cai[1], Rubao Ji[2]
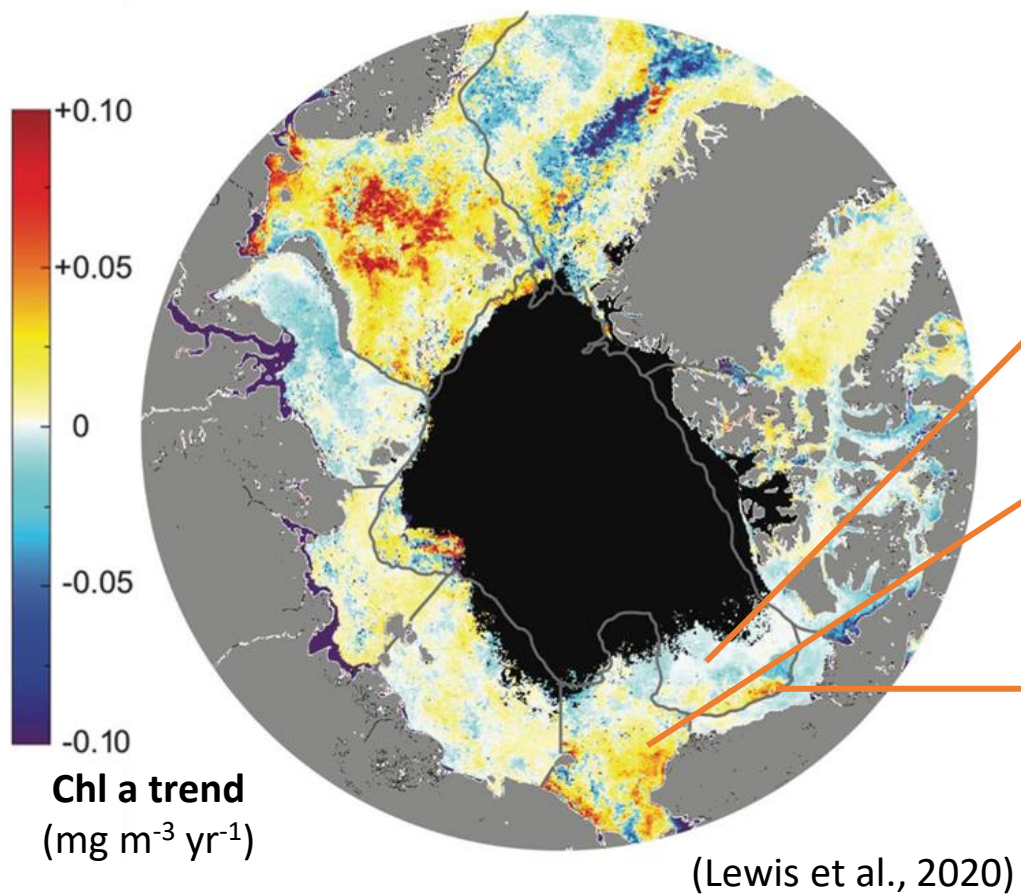
1. University of Delaware
2. Biology Department, WHOI

Outline:

- **Background**
- **Data for Modeling** (what data do we need? How do we use data?)
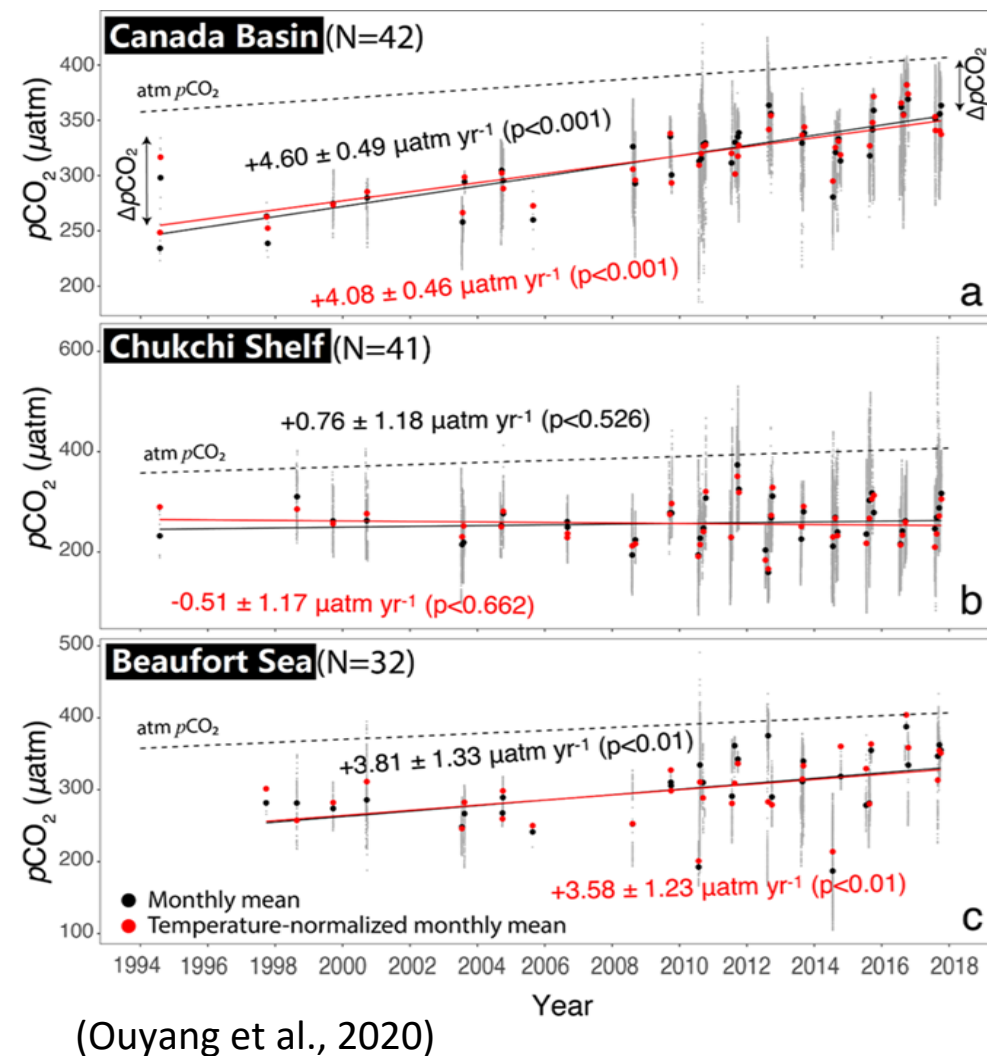- **Case Studies**
- **Synthesis Papers**

Significant long-term increase of Chl a
in most part of the western AO

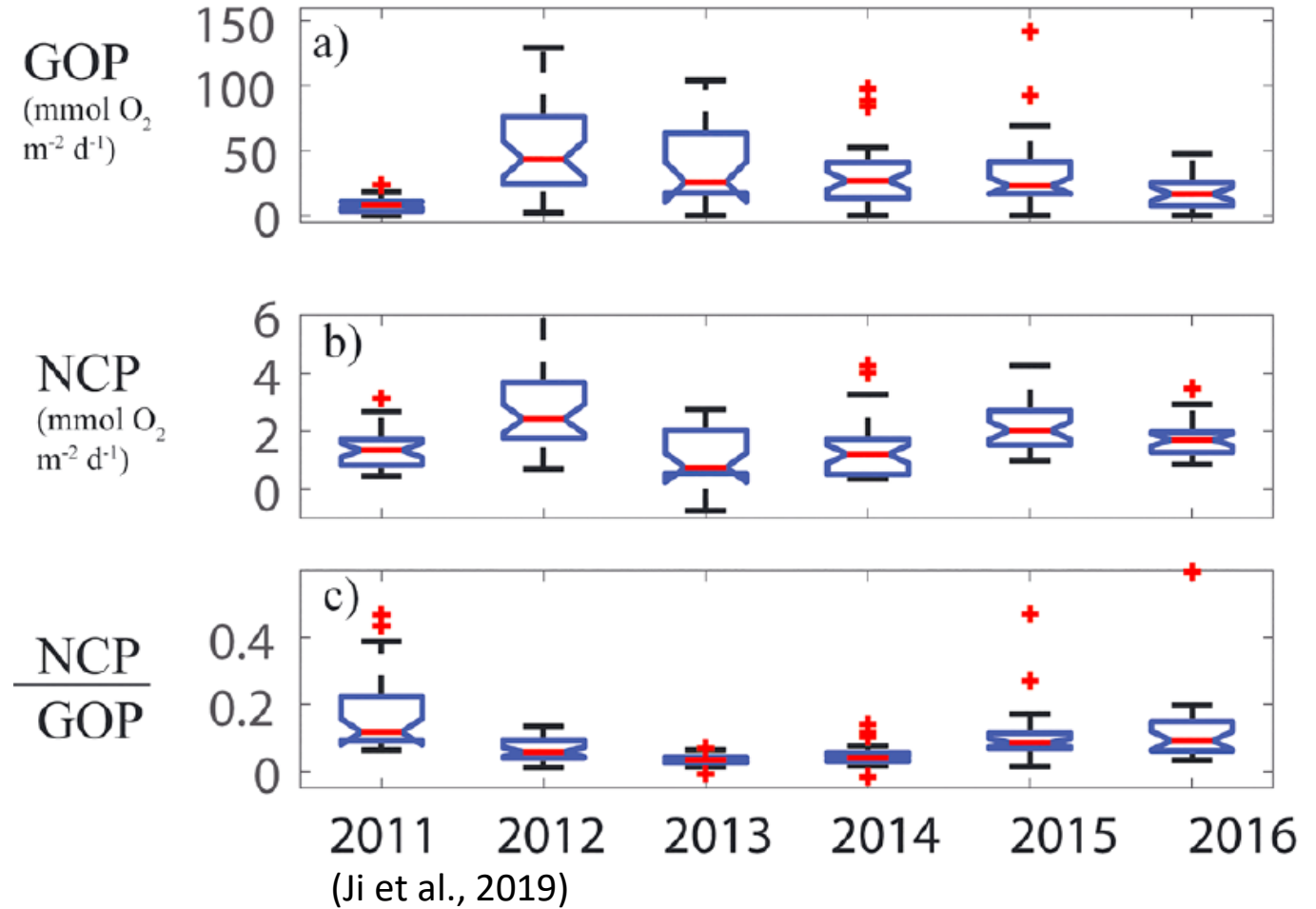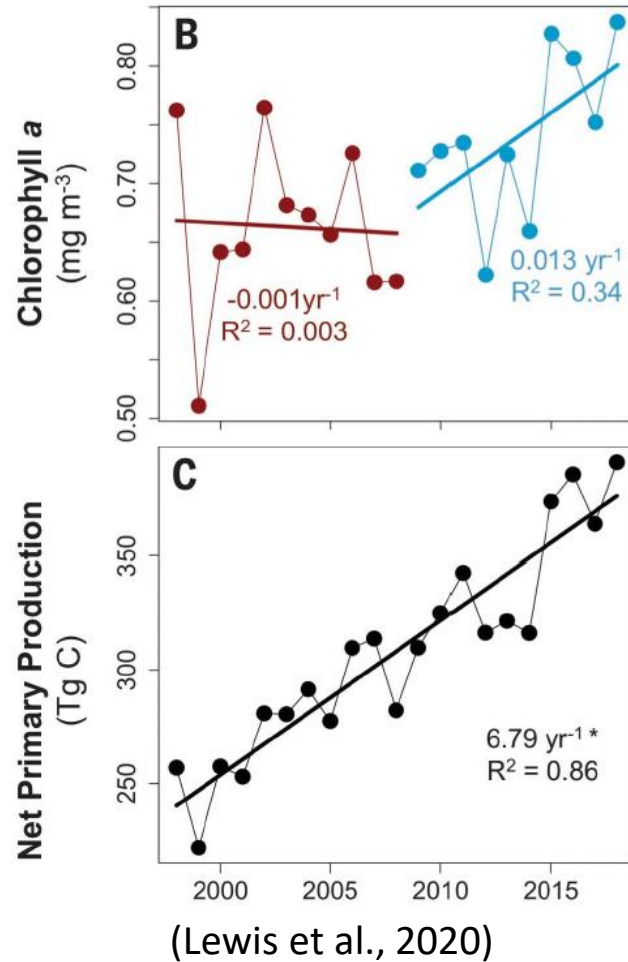These physical and biological changes lead to
heterogeneous $p\mathrm{CO_2}$ responses

Chl a trend
(mg m^{-3} yr^{-1})

(Lewis et al., 2020)

(Ouyang et al., 2020)

Background (cont'd)

- Net community production (NCP) is not linearly related to biological production rates.
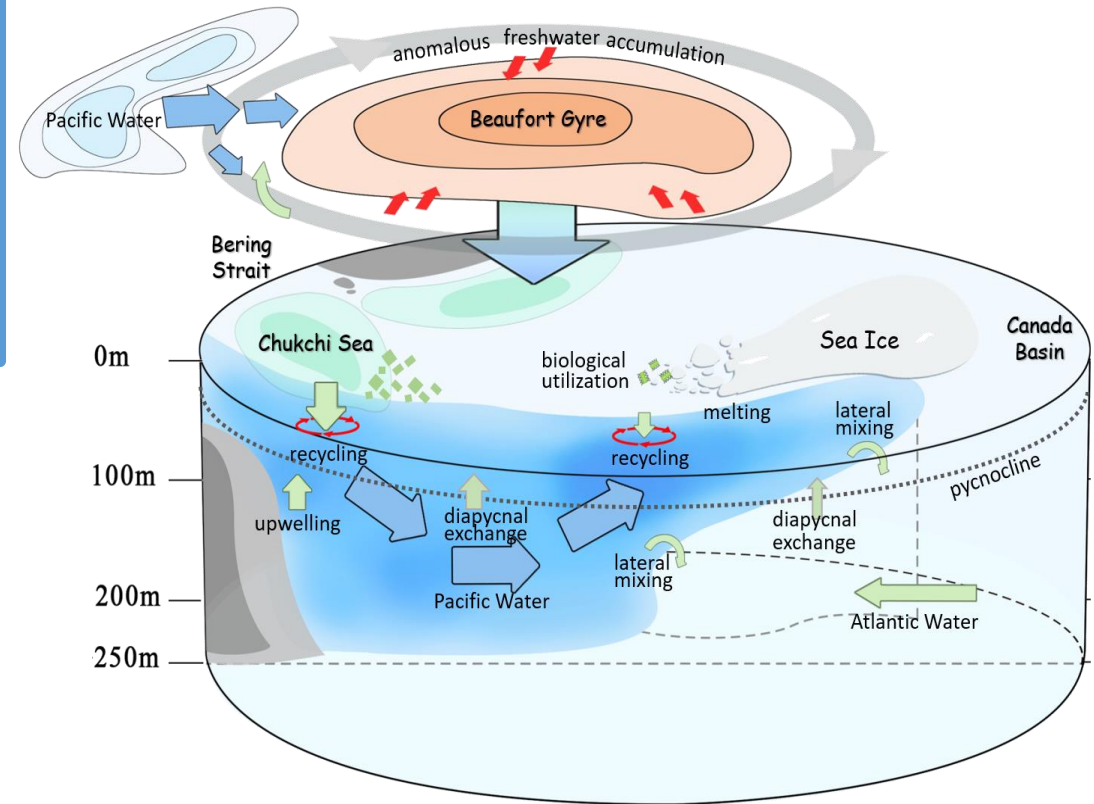

(Lewis et al., 2020)

(Ji et al., 2019)

3

What data do we use? (examples)

- Sea ice concentration
- Temperature
- Salinity
- Wind speed and direction

- NO_3
- pCO_2
- DIC
- pH
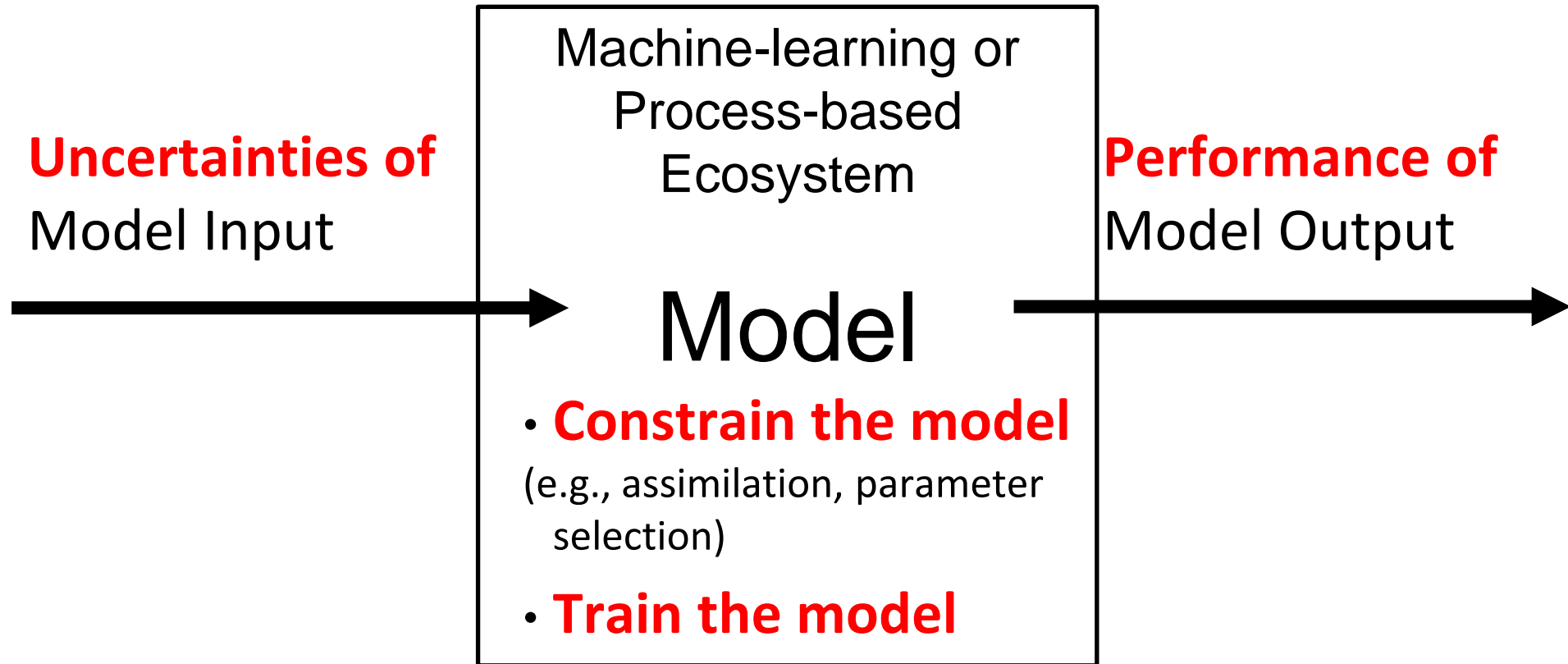
- Chlorophyll
- Gross Primary Production (GPP)
- O_2/Ar

and more …

anomalous freshwater accumulation

Pacific Water

Beaufort Gyre

Bering Strait

Canada Basin

Chukchi Sea

biological utilization

Sea Ice

melting

lateral mixing

0m

recycling

recycling

pycnocline

100m

upwelling

diapycnal exchange

diapycnal exchange

lateral mixing

200m

Pacific Water

Atlantic Water

250m

modified after Qi et al., (2017)

How do we use data?

Uncertainties of
Model Input

Machine-learning or
Process-based
Ecosystem

Model

- **Constrain the model**
(e.g., assimilation, parameter
selection)

- **Train the model**
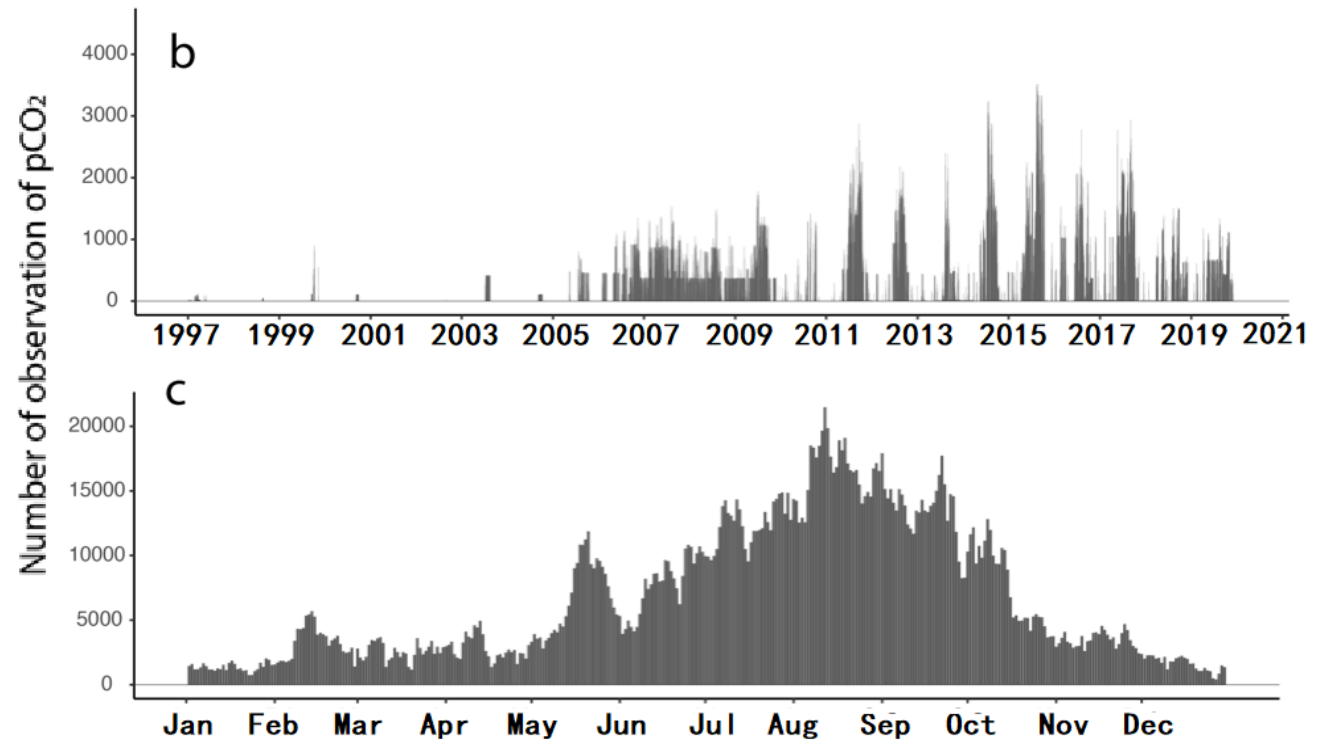
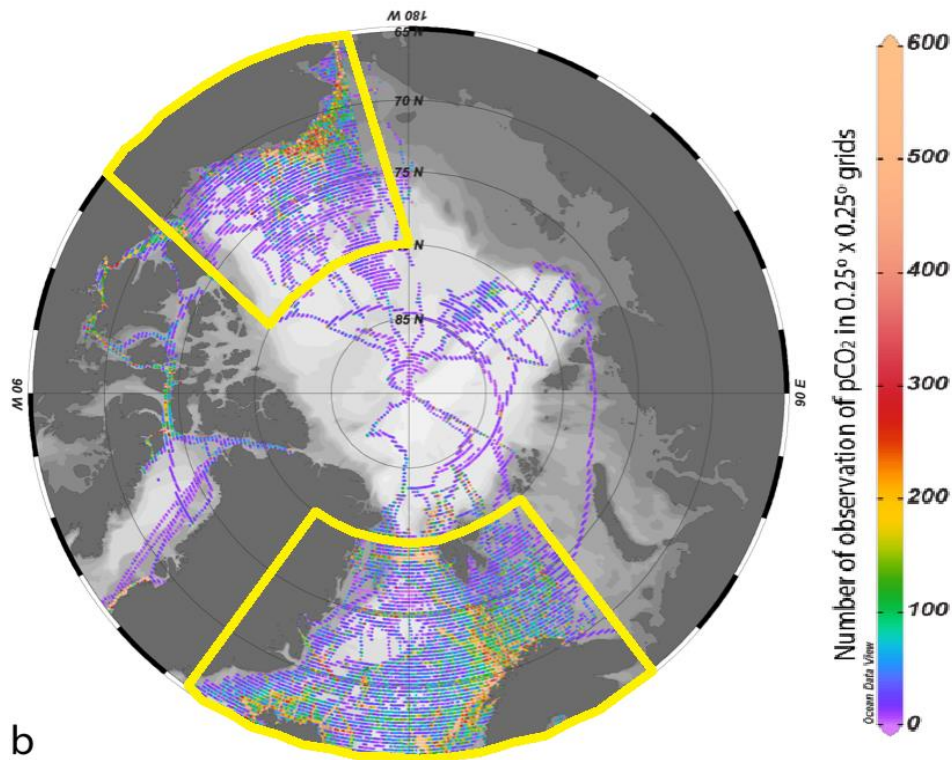**Performance of**
Model Output

SOCAT
SURFACE OCEAN CO₂ ATLAS
www.socat.info

Compiled pCO_2 data

- Most data are available in the Pacific and Atlantic sectors

- Data in the high-latitude Arctic Ocean remain sparse

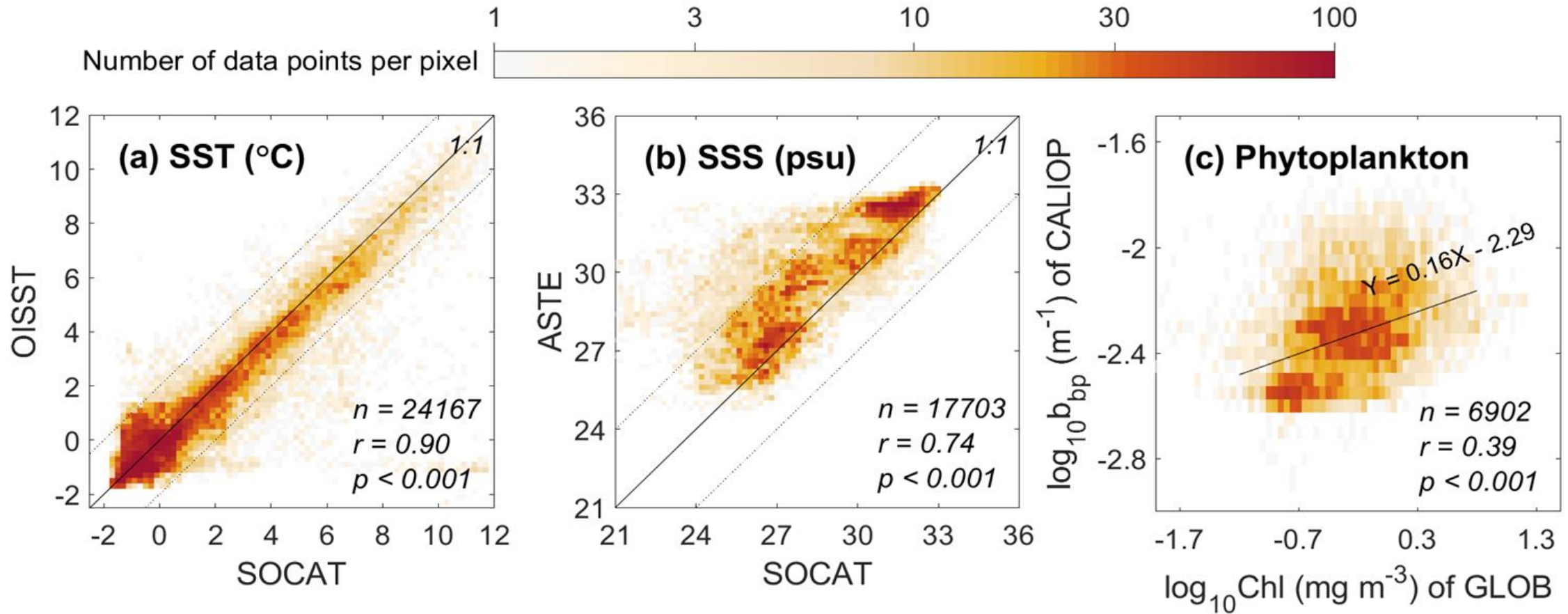- Recent data of 2014-2020 nearly doubled the total number

Compiled O$_2$/Ar data

- Concentrated in Chukchi and Beaufort Seas and few in high-latitude open-water area
- Available in most years from 2011-2021 and from late spring to early winter
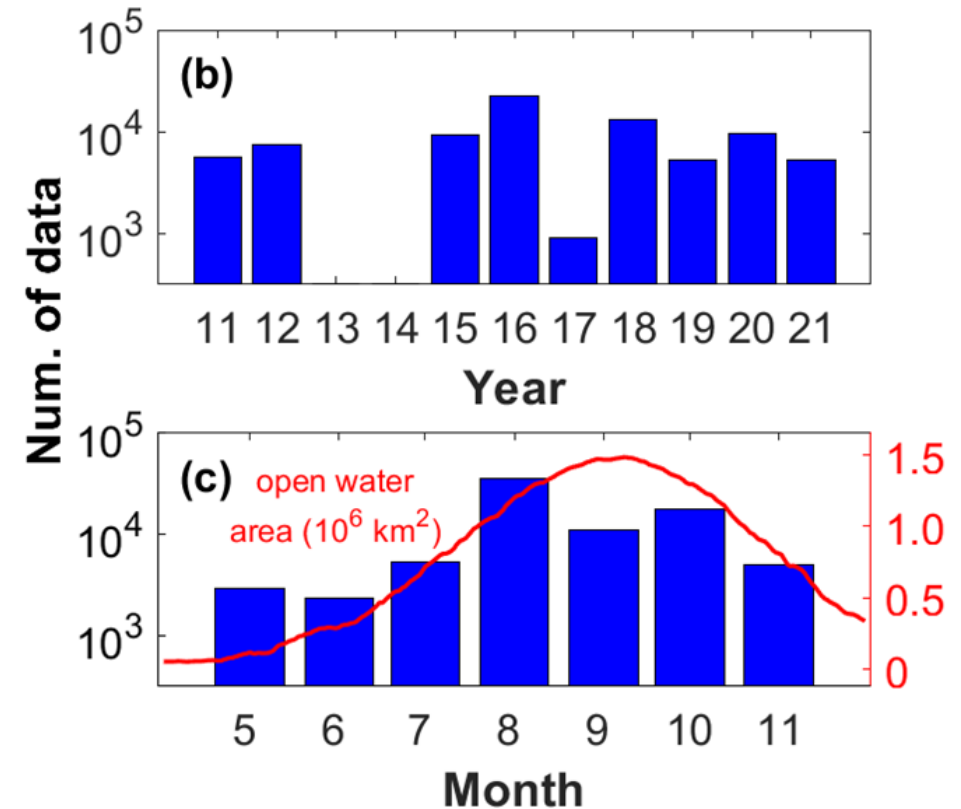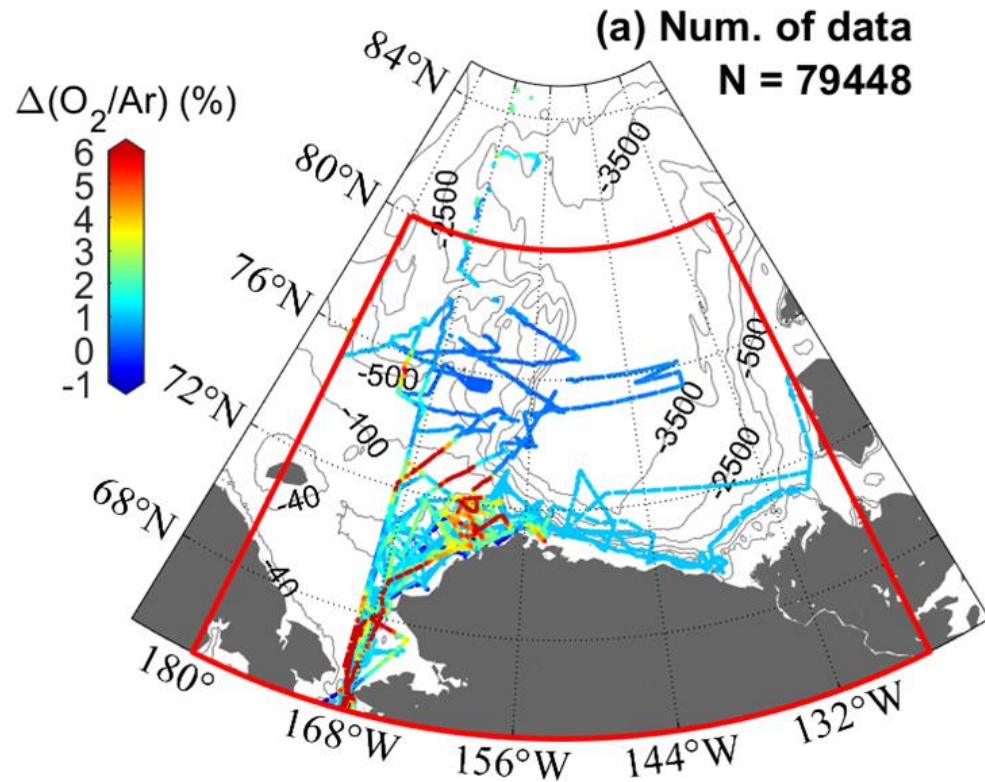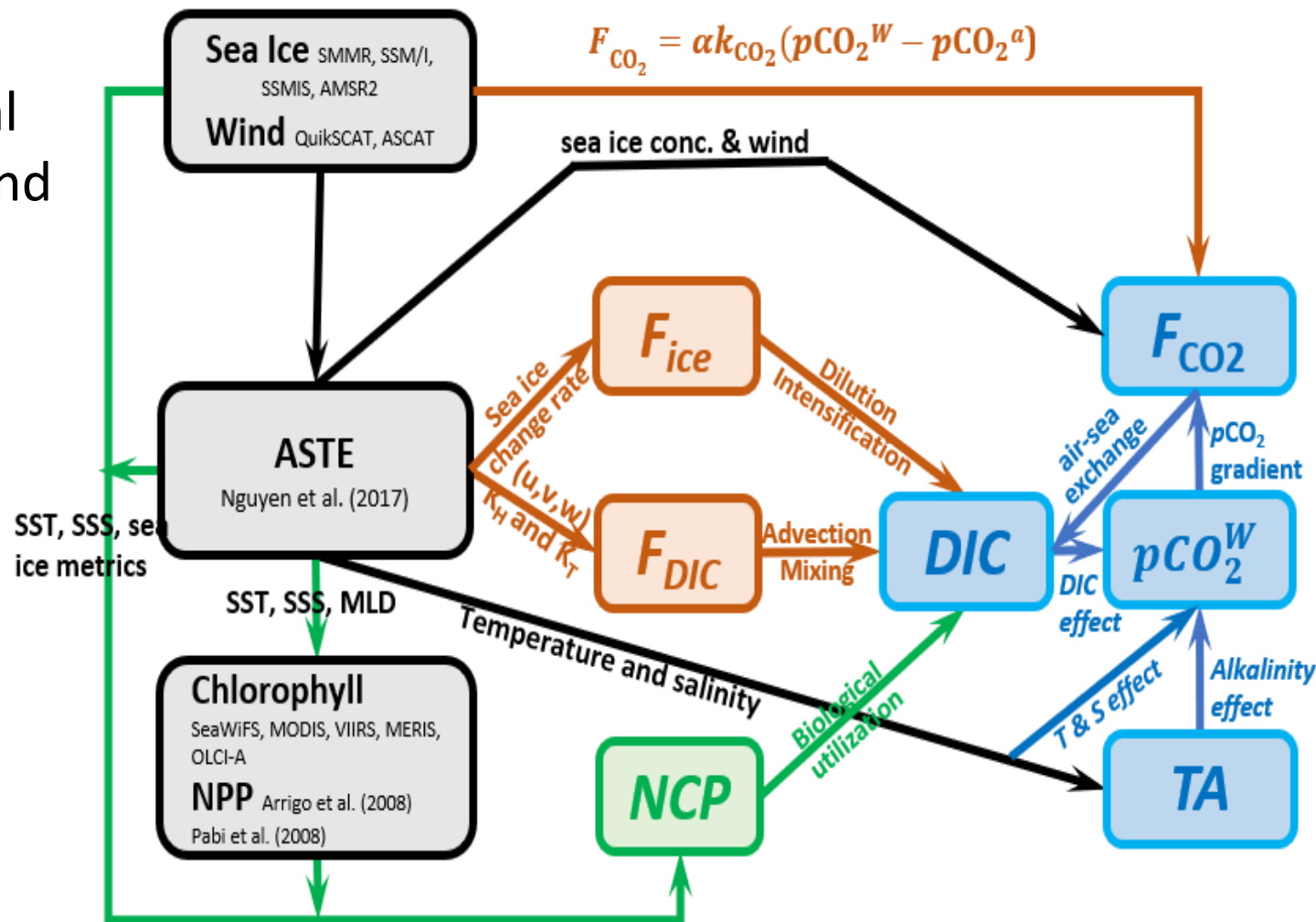
Uncertainties of Model Input

- Uncertainties vary across variables

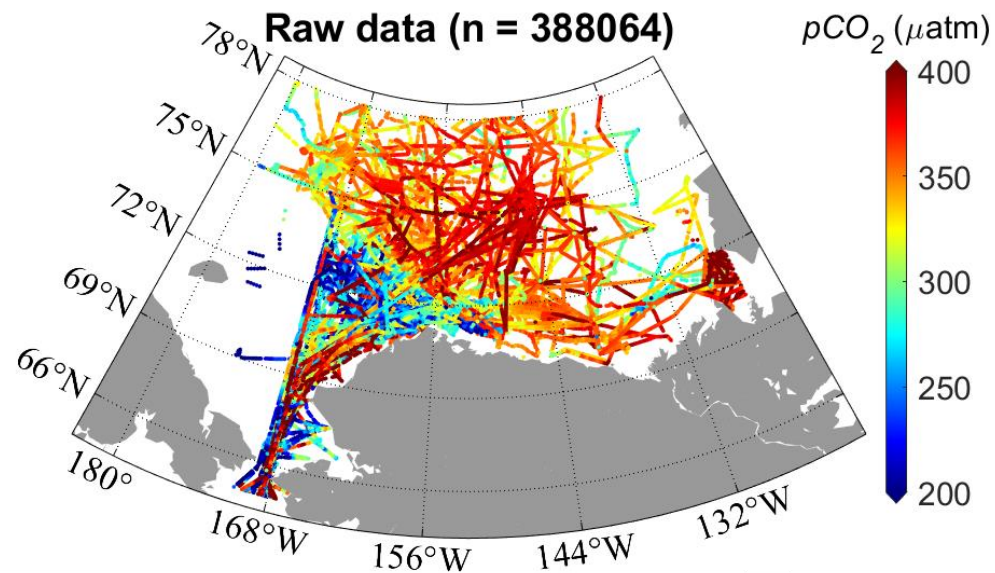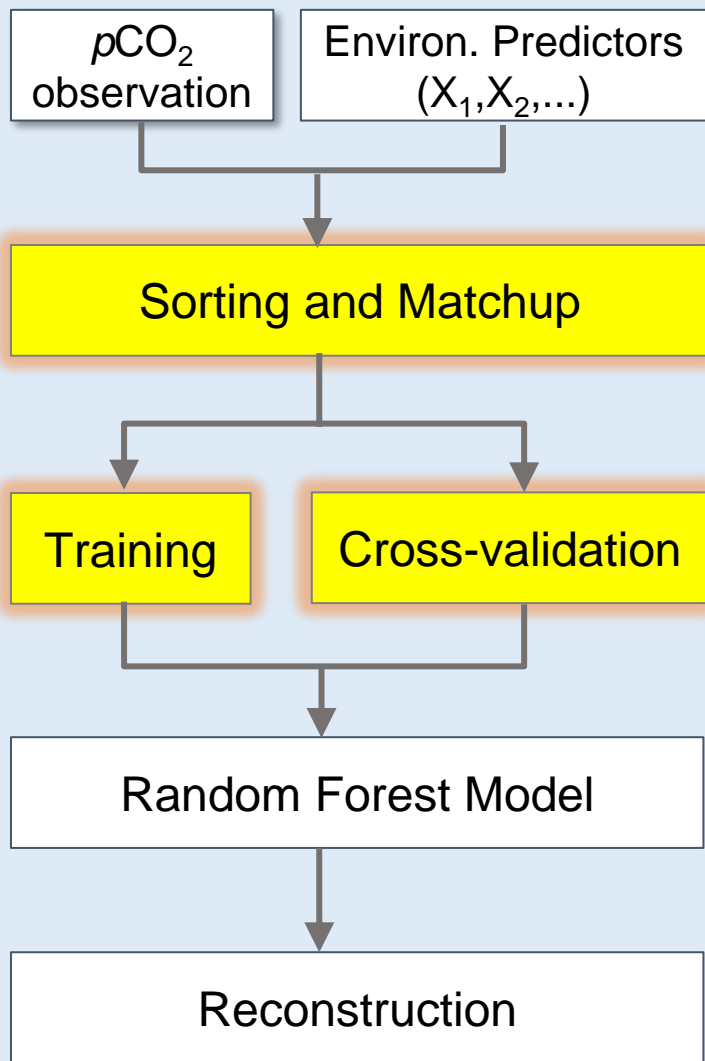- Discrepancies are location- and season-dependent.
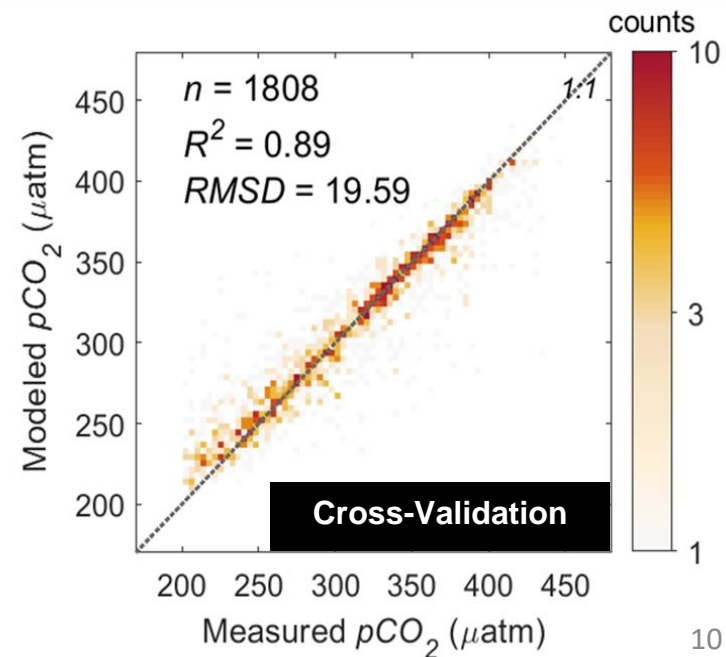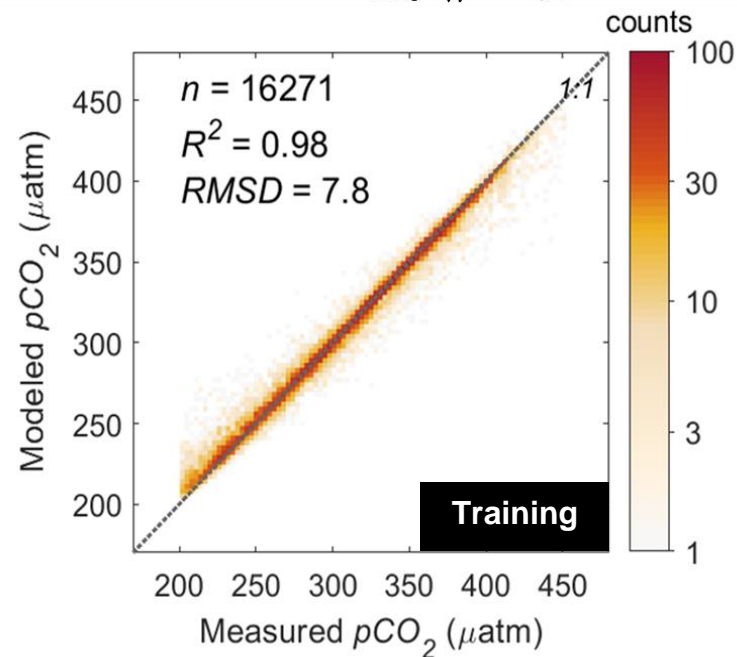
Constrain Process-based Model

- Constrain rates related to physical exchange, biological utilization, and chemical reaction processes.
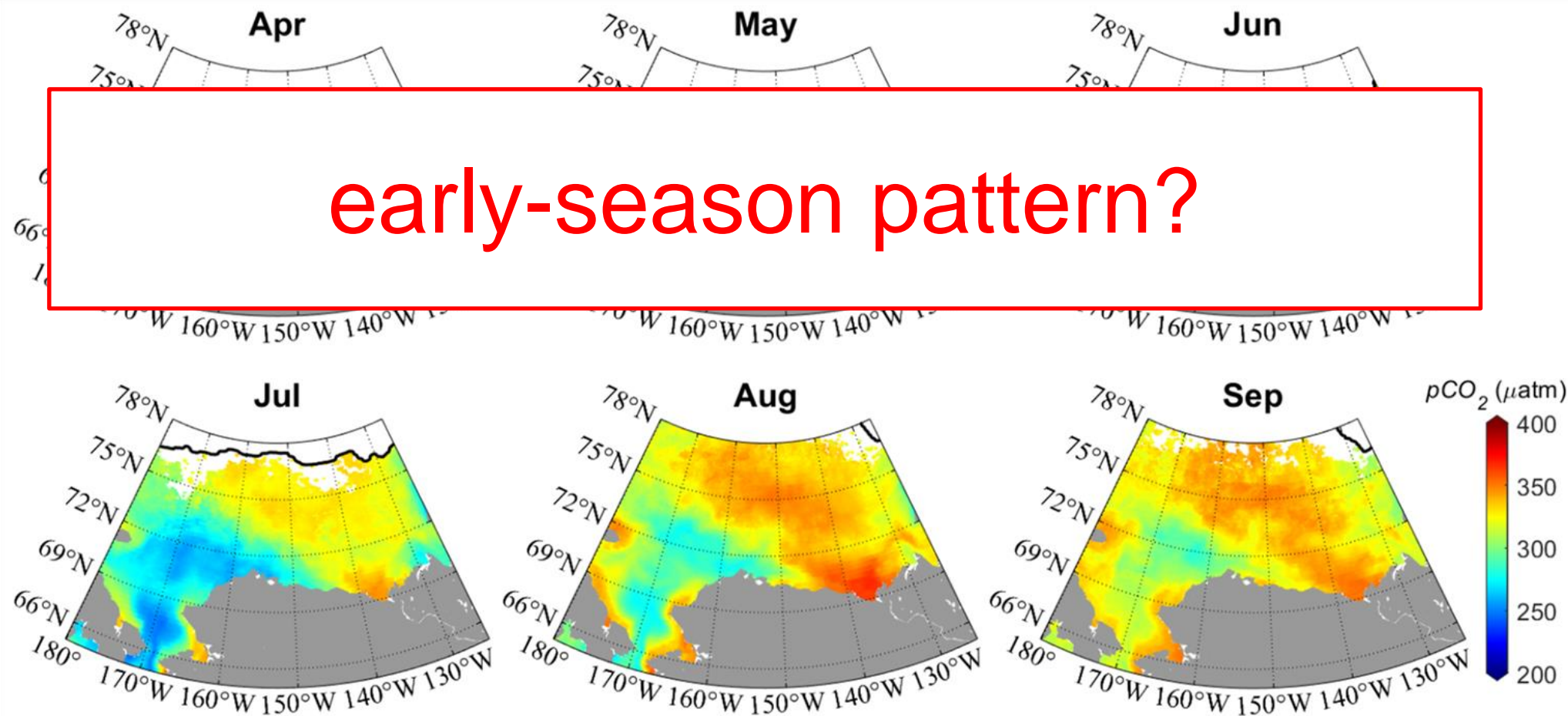
Train Process-based Model

- About 0.4 million pCO_2 measurements in the past decade

- Strong shelf-to-basin gradient

Flowchart:
- pCO_2 observation
- Environ. Predictors $(X_1, X_2,...)$
- Sorting and Matchup
- Training
- Cross-validation
- Random Forest Model
- Reconstruction

Raw data (n = 388064)

pCO_2 (μatm)

Training: $n = 16271$, $R^2 = 0.98$, $RMSD = 7.8$

Cross-Validation: $n = 1808$, $R^2 = 0.89$, $RMSD = 19.59$

Measured pCO_2 (μatm)

Modeled pCO_2 (μatm)

Early season during ice retreat (Apr to Jun)?

early-season pattern?

Early season during ice retreat (Apr to Jun)

- reversed spatial gradients at low latitudes
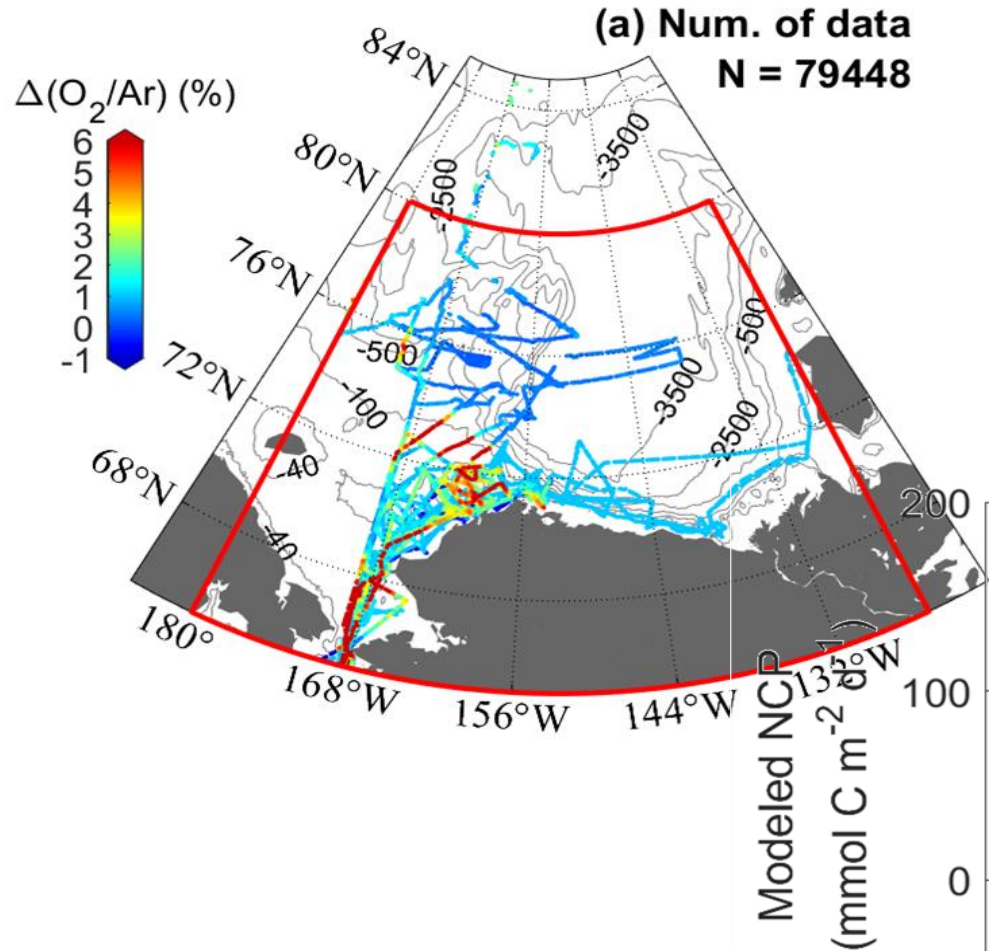- high pCO_2 in the newly opened water along sea ice edge
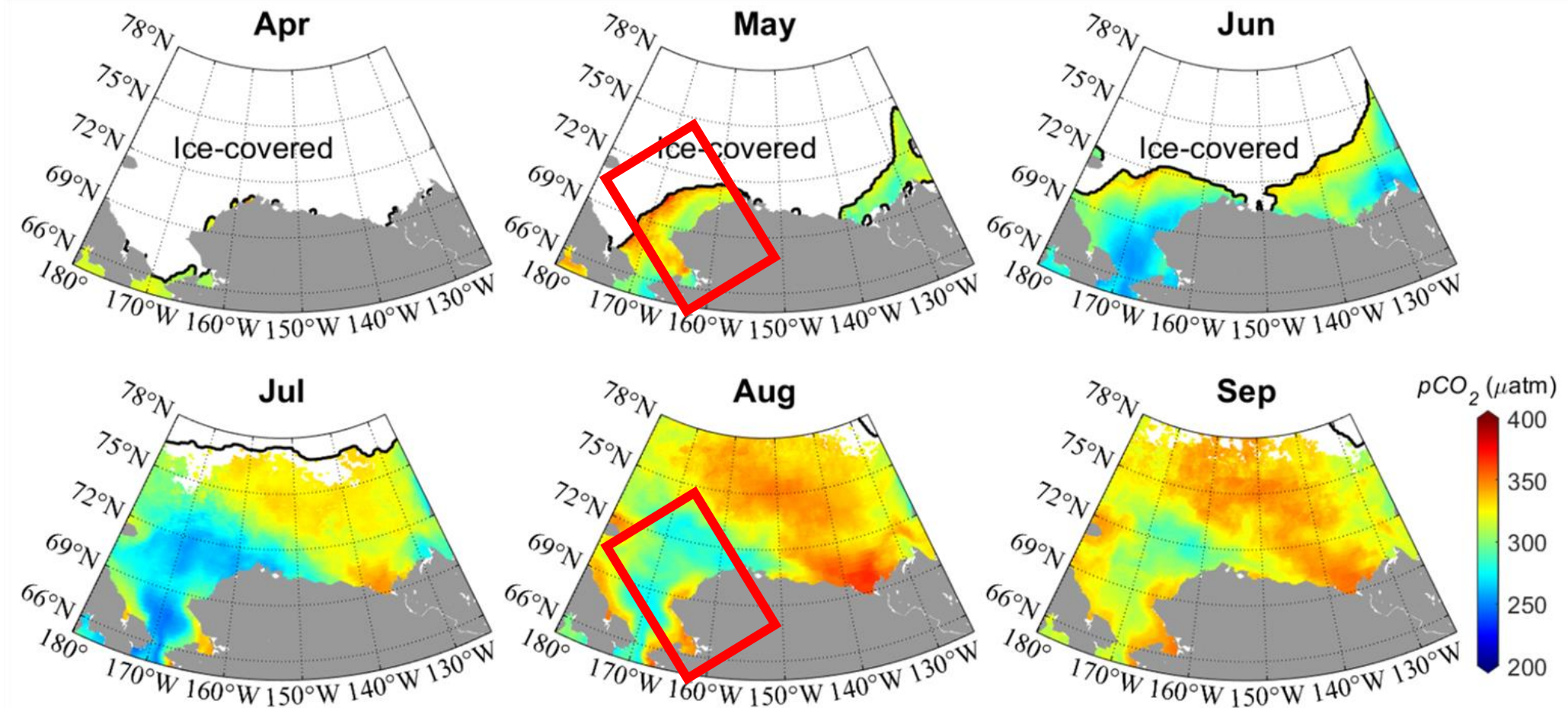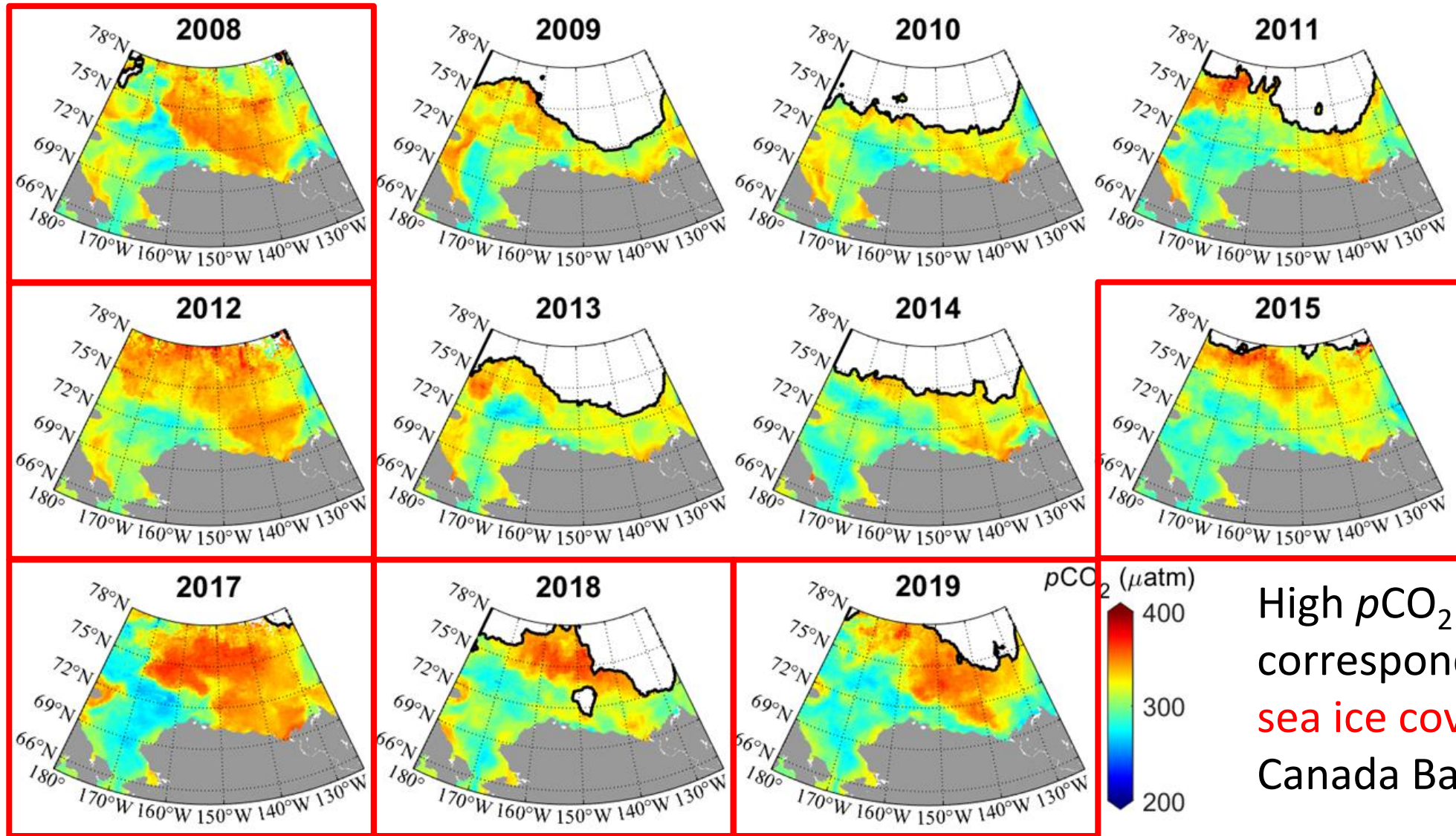
High pCO_2 generally corresponds to low sea ice cover in the Canada Basin

(a) Num. of data
N = 79448

$\Delta(O_2/Ar)$ (%)

Random Forest Model:

$$NCP = f(SSS, Daylength, SST, wind^2, Chl, SIC)$$

(a) Training

1:1

n = 4187
R^2 = 0.98
p < 0.001
RMSD = 2.14

(b) Cross-valid

1:1

n = 739
R^2 = 0.90
p < 0.001
RMSD = 5.32

Modeled NCP (mmol C m^{-2} d^{-1})

Observed NCP (mmol C m^{-2} d^{-1})

14
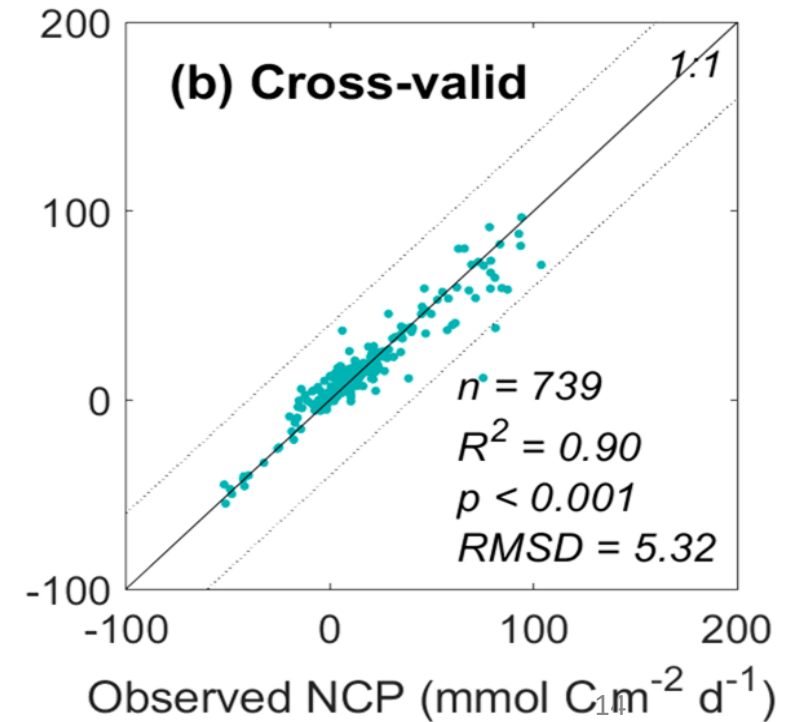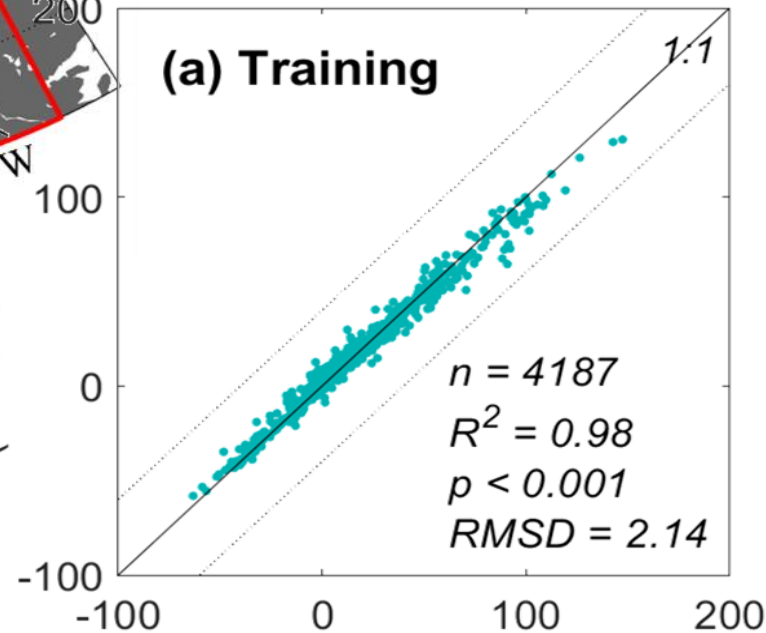
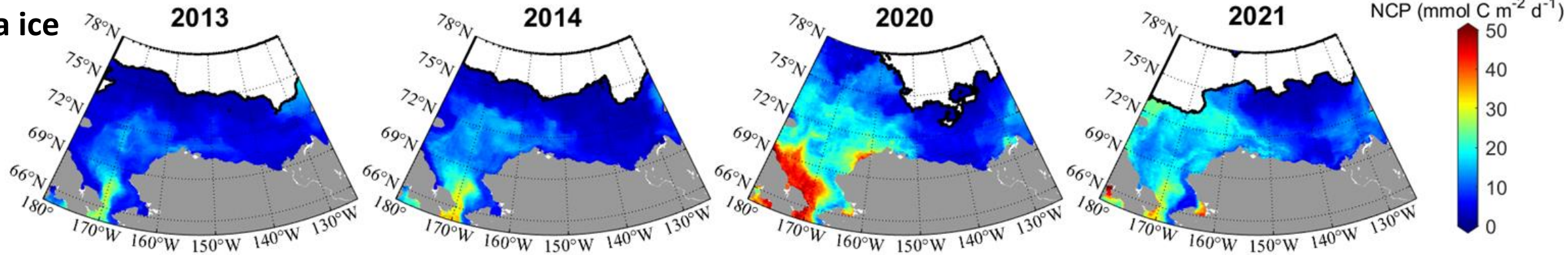**ML models produce weekly distribution of Net Community Production in the Western AO.**

Low sea ice cover:
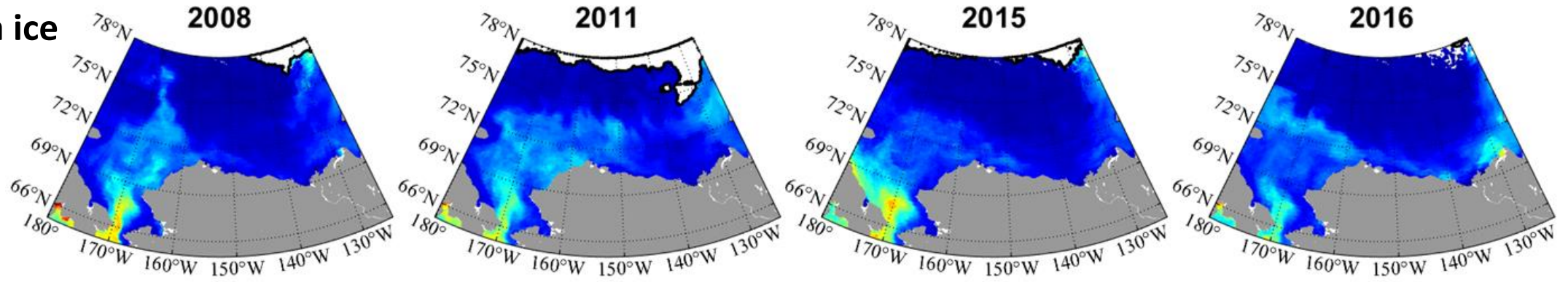
High sea ice cover:

Motivation:

Models and approaches use underlying assumption for dynamic processes to estimate the carbon sink in the Arctic Ocean:

- Ecosystem Model: 153±14 TgC yr^{-1} (Manizza et al., 2019)
- Atmospheric Inversion Model: 400~600 TgC yr^{-1} (Wanninkhof et al., 2013)

Objectives: *A **data-based** carbon flux and budget analysis*

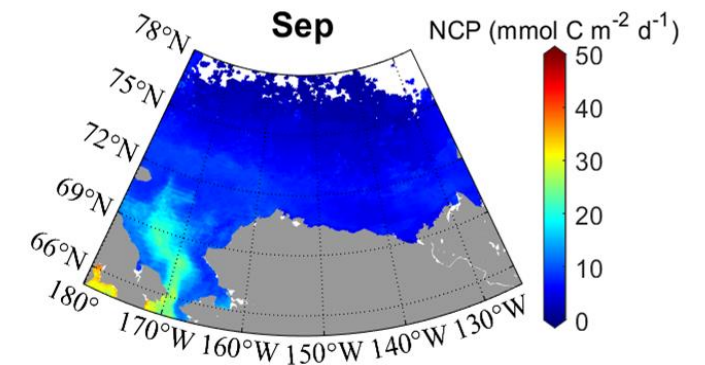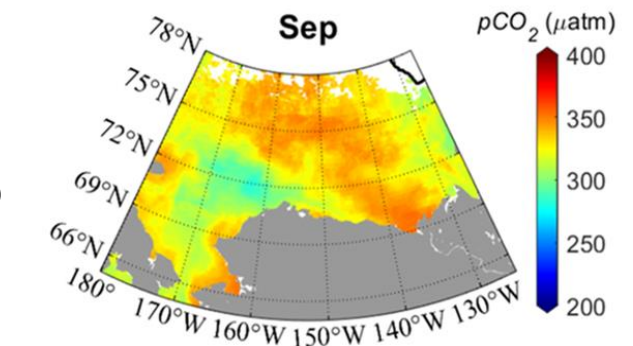**O1:** Provide independent carbon sink estimation

O2: Quantify the relative contribution of NCP and air-sea fluxes to carbon budget

$$F_{CO_2} = K_s \cdot k_{CO_2} \cdot (1 - SIC) \cdot \Delta pCO_2$$

Case Study I

$$\frac{d \int_0^{MLD} DIC dz}{dt} = F_{CO_2} + F_{ice} + K \cdot \Delta DIC_{vert} + NCP$$

Case Study I

Case Study II

Objectives:

O1: Biological production regimes classification (e.g., recycle- vs export-dominant system)

O2: Identify key environmental drivers and dynamics of pCO$_2$ and NCP seasonality

O3: Detect seasonal timing and duration and their impacts on year-to-year changes of pCO$_2$ and NCP