Data of Geology, Ecology, Oceanography, Space Science and Polar Science [4.3] : 15-23 (2022)
DOI https://dx.doi.org/10.22761/DJ2022..4.3.003

[Oceanography] Article

기계학습과 관측자료를 활용한 21년간의 로스해 표층 이산화탄소 분압 (fCO_2) 분포 재현

모아라 · 최정옥 · 박기홍*
극지연구소 해양연구본부, 인천시 연수구 송도미래로 26, 21990, 대한민국

Reconstruction of monthly fCO_2 distribution in the Ross Sea, Antarctica during 1998 -2018 using machine learning technique and observational data sets

Ahra Mo · Jung-Ok Choi and Keyhong Park*
Division of Ocean Sciences, Korea Polar Research Institute, Incheon 21990, Republic of Korea

Received: 27 September 2022, revised: 05 October 2022, accepted: 14 October 2022

요약문 해양은 인간활동에 의해 발생된 이산화탄소의 저장고이며, 특히 남극해는 인간기원의 이산화탄소의 약 40%를 흡수하는 해역으로 알려져 있다. 로스해는 남극해에서 가장 생산력이 높은 지역이나, 그 이산화탄소의 흡수력에 관해서는 아직 명확하지가 않다. 이는 남극의 특성상 관측의 시기와 지역의 제한이 주요한 요인이다. 본 연구에서는 이러한 관측기반 자료의 한계를 극복하기 위해 기계학습을 통해 표층 이산화탄소 분압(fugacity of Carbon dioxide; fCO_2)의 농도를 재현하였으며, 이를 위해 기존의 현장관측 자료 뿐만 아니라, 인공위성 및 모델 자료가 활용되었다. 또한, 수온, 해빙 농도, 클로로필 농도와 같은 해양환경 변수 이외에 운량과 풍속 그리고 엘니뇨 인덱스를 학습에 추가하여 더욱 정확한 fCO_2의 농도 재현을 위해 노력하였다. 재현은 기계학습의 한 종류인 랜덤 포레스트 기법을 사용하였으며, 이를 통해 인공위성을 통한 클로로필 농도 등이 제공되기 시작한 1998년부터 2018년까지의 지난 21년간 남극 로스해의 fCO_2의 시공간의 변동을 월별로 제공하고자 한다.

주요어: 로스해, 기계학습, 이산화탄소, 랜덤 포레스트

Abstract The ocean is a major reservoir of anthropogenic carbon dioxide, especially the Southern Ocean has been known to absorb 40% of the carbon dioxide emitted by human activity. The Ross Sea is one of the most productive regions in the Southern Ocean; however, its carbon dioxide absorption capacity has not been clearly evaluated yet. Because the Southern Ocean is geographically isolated from civilization and thus, its remoteness prevents making sufficient observations from proving reliable carbon dioxide sink strength estimates. Thus, in order to overcome the current spatial and temporal limitations of direct observations, the fugacity of carbon dioxide (fCO_2) data was reproduced using a machine learning technique (i.e., random forest technique). The

*Corresponding author: keyhongpark@kopri.re.kr

technique is a type of machine learning frequently used to reproduce marine environmental variations through training satellite data and modeled data as well as existing observational data. Furthermore, to reproduce more reliable fCO_2 estimates, in addition to marine environmental variables (i.e., sea surface temperature, sea ice concentration, and chlorophyll-a concentration), cloud cover, wind speed, and El Niño index were included in the machine learning procedure. In this study, we provide the past 21 years (1998 – 2018) of monthly spatial and temporal variation information of dissolved carbon dioxide in the Ross Sea, Antarctica.

Keywords: Ross Sea, machine learning, carbon dioxide, random forest

1. 서론

해양은 인간활동으로 발생된 이산화탄소를 흡수하는 탄소 저장고로서, 2021년 글로벌 탄소수지보고서(Global Carbon Budget)에 따르면 지난 10년동안 발생한 인간기원 이산화탄소의 약 30%가 해양으로 흡수된 것으로 보고되었다(Friedlingstein et al., 2022). 남극해의 면적은 전체 해양 면적의 약 20%를 차지하고 있지만, 전체 해양에서 흡수되는 인간기원 이산화탄소의 약 40%를 흡수하는 중요한 해역이다(Orr et al., 2001; DeVries 2014). 특히, 본 연구가 수행된 남극 로스해(Ross Sea) 해역은 대기중의 이산화탄소를 강하게 흡수하는 해역으로, 연간 약 7.5~13 Tg의 탄소가 흡수되는 것으로 보고된 바 있다(Arrigo et. al., 2008; DeJong and Dunbar, 2017). 해양에서 흡수되는 탄소의 양을 추정하기 위해서는 현장 관측이나 모델에서 모의된 표층 이산화탄소 분압 자료가 요구된다. 해양의 표층 이산화탄소 분압은 이산화탄소의 몰 분율(mole fraction; $x(CO_2)$)에 총압력을 곱한 pCO_2로 표현하거나, 실제 환경에서의 비이상적 거동을 고려한 퓨개시티(fugacity of CO_2; fCO_2)로 표현할 수 있다.

남극해는 지리적인 특수성으로 인해 타 해역에 비해 fCO_2(또는 pCO_2)의 현장 관측 자료가 부족한 해역으로, 특히 여름철을 제외한 나머지 계절의 현장 관측자료가 매우 부족하다(Fay et al., 2018). 이러한 현장 관측 자료의 부족은 모델을 통해 모의된 결과의 정확성을 감소시키는 주요 요인으로 작용한다. 최근 무인 관측 장비(무인선, 관측 부이 등)의 개발로 계절에 상관없이 지속적인 남극해 현장 관측 수행이 가능해 졌지만, 해류의 영향으로 관측 자료의 지역적인 편향이 생길 수 있다(Sutton et al., 2021; Williams et al., 2017). 현재까지 보고된 각 플랫폼 별 남극해 pCO_2 관측치(혹은 추정치)의 불확실도는 현장 관측과 관측 부이에서 각각 2 μatm, 11 μatm으로 보고되었다(Bakker et al., 2016; Williams et al., 2017).

Breiman (2001)의 연구에서 제시된 랜덤 포레스트는 기계학습의 한 종류로, 다수의 결정 트리(decision tree)들을 학습하는 앙상블 방법이다. 이 알고리즘은 트리 수와 학습집합의 크기 등에 의해 랜덤 포레스트의 성능과 예측 결과의 정확성이 결정되기 때문에 최적의 트리 수와 학습집합의 크기를 결정하는 것이 중요하다(Ma & Fan, 2017). 현재까지 이 알고리즘은 과학 분야를 포함한 모든 연구분야에서 널리 활용되고 있다. 그 중 현장 관측 자료와 인공위성 자료를 랜덤 포레스트에 학습시켜 멕시코 만의 표층 pCO_2를 추정한 Chen et al. (2019)의 연구에서 예측 결과의 불확실도는 10 μatm 이하로 보고되었다. 또한 랜덤 포레스트와 신경망 모델(Self-organizing maps-feed forward network; SOM-FFN)을 사용하여 추정된 남극해 pCO_2의 평균 제곱근 편차가 각각 14.84 μatm, 16.45 μatm으로, 랜덤 포레스트의 예측 결과가 SOM-FFN의 그것과 크게 다르지 않음이 확인되었다(Gregor et al., 2017).

본 연구에서는 현장 관측자료와 위성 자료를 전 처리하여 랜덤 포레스트 알고리즘을 학습 시키고, 결과를 도출하는 방법에 대해 간략하게 다루고 있다. 또한, 이에 따른 결과로 1998년부터 2018년 동안의 남극 로스해(55°S~72°S, 145°E~145°W) 표층 fCO_2의 월별 분포를 모의하였으며, 추정된 데이터 자료를 제공하고자 한다.

2. 본론

2.1. 격자데이터

본 연구에서 사용된 데이터는 주로 원격감지데이터를 사용하였다(Table 1). 표층 혼합층의 깊이(mixed layer depth; MLD)는 동화모델(GLORYS12V1; Ferry et al., 2010)의 산출물을 사용하였으며, 데이터의 시간적 범위는 1998년부터 2018년까지이다. 데이터의 공간분해능은 동일한 0.25°로 재격자화하였으며, 시간분해능은 모델 훈련 및 검증의 경우엔 8 days, 모델 예측의 경우엔 1 month로 평균을 취하였다. 또한, 모델 훈련을 위해 SOCAT v3 fCO_2 자료가 사용되었다(Bakker et al., 2016).

Table 1. Data used in this study. The temporal and spatial resolutions are for the raw data (before gridding).

Variables	Description	Unit	Resolution		Source & Reference
			Time	Space	
fCO_2	Fugacity of carbon dioxide	μatm	1 month	1°	http://www.socat.info/ Bakker et al. (2016)
Chl-a	Chlorophyll a concentration	$mg{\cdot}m^{-3}$	8 days	25 km	http://www.globcolour.info/ Maritorena and Siegel (2005)
SST	Sea surface temperature	kelvin	1 day	0.25°	http://www.ncdc.noaa.gov Reynolds et al. (2009)
SIC	Sea ice concentration	1	1 day	25 km	http://nsidc.org/ Comiso and Nishio (2008)
MLD	Mixed layer depth	m	1 day	1/12°	http://www.mercator-ocean.fr Ferry et al. (2010)
TCC	Total cloud cover	(0 – 1)	1day	25 km	https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5 Hersbach et al. (2017)
Wind	Wind speed	$m{\cdot}s^{-1}$	1day	25 km	https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5 Hersbach et al. (2017)
ONI	Oceanic Nino index	-	1 month	-	https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_v5.php Huang (2017)

2.2. 입력데이터의 변환과 변수추가

모델 개발에 앞서 원시 데이터의 누락된 값을 처리하고 데이터를 변환하여 얻은 데이터셋을 모델의 입력변수(feature)로 사용하였다. 자료의 비대칭도(skewness)가 큰 MLD와 클로로필(Chl-a)은 정규분포에 가까워지도록 상용로그 변환을 하였다(Nakaoka et al., 2013; Landshuster et al., 2014; Gregor et al., 2017). 자료의 계절변화를 표현하기 위하여 날짜 순서(day of year)를 변수로 활용한 연구(Zeng et al., 2014; Gregor et al. 2017)가 있으나 본 연구에서는 월(month)의 사인함수(sin_t)와 코사인함수(cos_t)를 추가변수로 활용하였다(**Eq.1**& **Eq.2**).

$$\cos\_t = \cos\left(\frac{2\pi \times \text{month}}{12}\right) \quad \textbf{(Eq.1)}$$

$$\sin\_t = \sin\left(\frac{2\pi \times \text{month}}{12}\right) \quad \textbf{(Eq.2)}$$

최종적인 입력변수는 총 구름의 양(total cloud cover; TCC), 해빙 농도(sea ice concentration; SIC), 표층수온(sea surface temperature; SST), 풍속(Wind), 해양 Nino 인덱스(oceanic Nino index; ONI), cos_t, sin_t, $\log_{10}$ (Chl-a), $\log_{10}$ (MLD)이며, 출력변수(label)는 fCO_2이다. 이 데이터를 분할하여 80%는 모델의 훈련에, 20%는 모델 테스트에 이용하였다. 본 연구에서는 난수의 시드(random seed)를 달리하여 10가지 서로 다른 분할된 훈련-테스트 데이터셋을 준비하였다.

2.3. 랜덤 포레스트 회귀

배깅(bagging)은 부트스트랩 집합체(bootstrap aggregation)로부터 온 말이다. 랜덤 포레스트는 의사결정나무(decision tree)에 배깅(bagging)을 적용한 모델이다(Breiman, 2001). 개별나무의 예측 값의 평균을 취하여 실제 값을 예측한다. 데이터 탐색과 모델 개발은 주로 파이썬(Python) 언어를 사용하여 수행하였다. Scikit-learn 패키지와 함께 여러 다른 add-on 패키지를 이용하여 모델 구축 및 튜닝 그리고 정확도 평가를 하였다. 랜덤 포레스트 회귀모델구축을 위해 의사결정나무의 개수는 500으로 설정하였으며, 랜덤 포레스트의 다른 파라미터들은 파이썬에서 제공되는 RandomForestRegressor 함수의 기본 설정 값을 사용하였다. 또한, 최적의 분할을 위해 고려할 최대 입력변수의 개수를 전체특성의 수로 하여 모든 특성을 고려하도록 설정하였지만, 동시에 bootstrap sampling을 하도록 설정하였기 때문에 여전히 무작위성은 존재한다.

3. 결과

3.1. SOCAT 데이터베이스의 이산화탄소 관측 자료 분포 분석

연구기간 동안(1998년 - 2018년) 로스해에서 관측된 이산화탄소 격자자료의 분포를 Fig 1.에서 보였다. 21년간 관측된 자료임에도 불구하고 남반구인 로스해의 겨울과 봄철 동안 관측 자료가 매우 부족하였으며(6월 1회, 7월 1회, 8월 1회, 10월 1회), 특히 9월은 관측자료가 전혀 존재하지 않았다. 이는 남극해가 동계기간 동안 해빙으로 덮여 있었기 때문이라 추정되며, 따라서 이 기간 동안의 추정결과의 신뢰도나 재현 결과의 정확도는 활용하기 어렵다고 판단된다. 하지만 남극 이산화탄소의 흡수는 주로 생산력이 집중되는 여름철에 발생하기 때문에, 남극해와 로스해의 이산화탄소의 변동 이해에는 동계기간 자료의 영향이 극히 제한적일 것으로 생각된다.

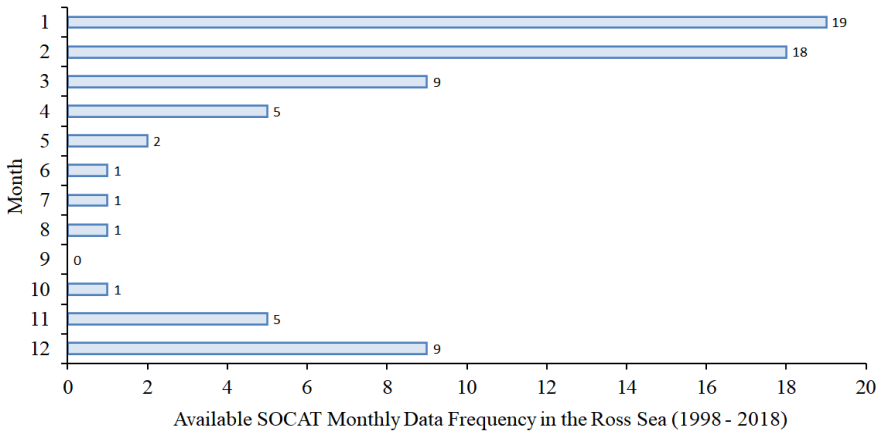

Figure 1. Available SOCAT monthly data frequency in the Ross Sea

3.2. 학습된 랜덤 포레스트 모델의 재현 성능 평가

Fig. 2에서는 랜덤 포레스트 모델의 성능의 평가를 위해 원시 데이터에서 추출한 fCO_2(x 축)와 기계학습모델을 통해 예측된 fCO_2(y 축)을 비교하였다. 총 10회의 반복실험을 통해 성능을 테스트하였고, 이의 결정계수(r^2)는 0.95이며, 평균제곱근편차(RMSE)는 12.3 μatm으로 나타났다.

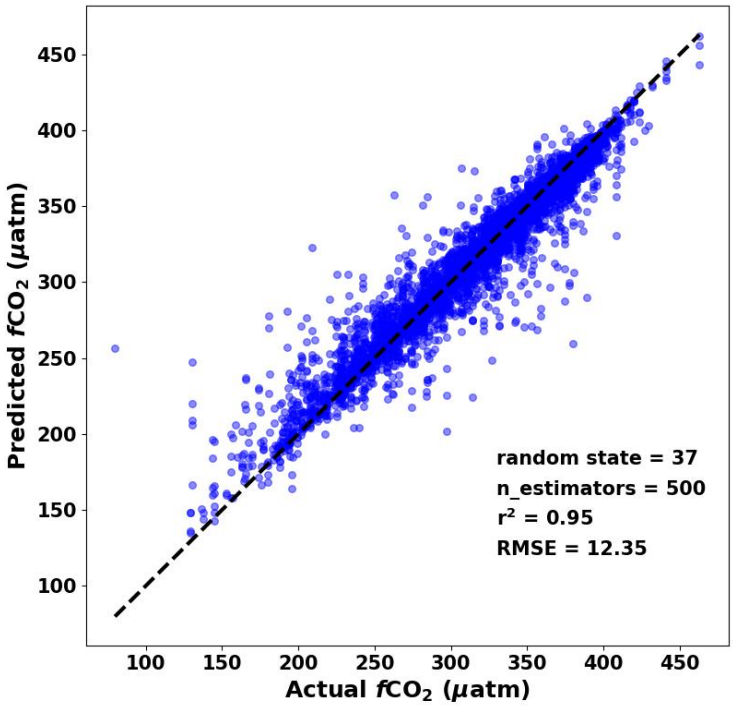

Figure 2. Evaluation of random forest performance. X-axis is sampled true fCO_2 values of original data set and y-axis is random forest-derived fCO_2 values. The dotted line is 1-to-1 line.

3.3 로스해 이산화탄소 분압(fCO_2) 재현 결과

랜덤 포레스트 기법을 통해서 남극 로스해의 fCO_2의 재현에 활용된 변수의 기여도 평가 결과, 클로로필의 농도가 fCO_2 결정에 가장 중요한 요인으로 이용되고 있음을 알 수 있었다(Fig. 3). 이는 생물 생산에 의한 이산화탄소의 흡수가 로스해에 중요한 역할을 하고 있다는 것을 암시한다. 그 외의 대부분의 변수들의 중요도는 큰 차이를 보이지 않았지만, 모두 유의미하게 활용 되었으며, 해빙의 역할은 상대적으로 이산화탄소의 흡수력 재현에 적게 기여하고 있음을 알 수 있었다.

Fig. 4에서는 랜덤 포레스트를 통해 추정한 fCO_2의 여름철(12월-3월)의 평균값을 도시하였다. 이를 통해 로스해의 fCO_2는 폴리냐가 열리는 12월과 1월 고위도(남부)에서 생산력의 증가와 함께 급격히 낮아짐을 알 수 있었다. 2월부터는 서서히 fCO_2가 증가하며, 3월부터는 해빙의 증가와 함께 지속적으로 증가하게 됨을 알 수 있었다. 특징적으로 2월부터 3월까지 장보고 기지가 위치한 로스해의 서편에서는 지속적으로 강한 이산화탄소의 흡수가 일어나고 있음을 볼 수 있었다. 또한, 기존 현장관측 자료가 전혀 없는 9월의 해빙이 없는 로스해 북부 해역에서 본 연구에서 사용된 입력변수들(Table 1 & equations 1–2)의 격자 데이터를 획득할 수 있었기 때문에, 이 기간의 표층 fCO_2 또한 본 연구를 통해 재현되었다(supplemental material). 하지만, 이 기간동안 재현된 fCO_2는 낮은 신뢰도를 가질 것으로 예상되기 때문에 자료를 활용하는데 어려움이 있을 것으로 판단된다.

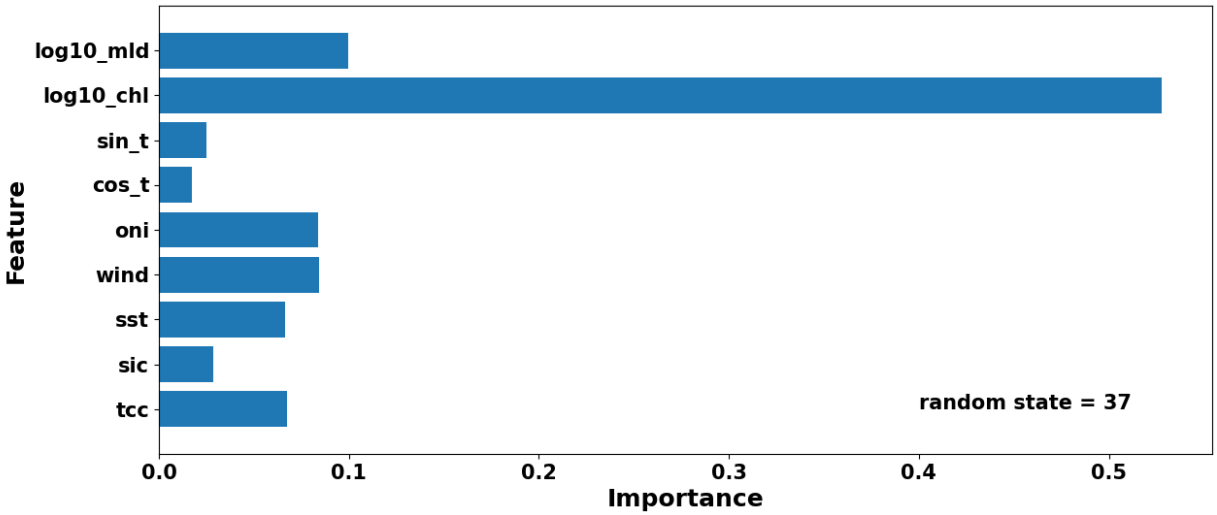

Figure 3. Feature importance of fCO_2 estimate

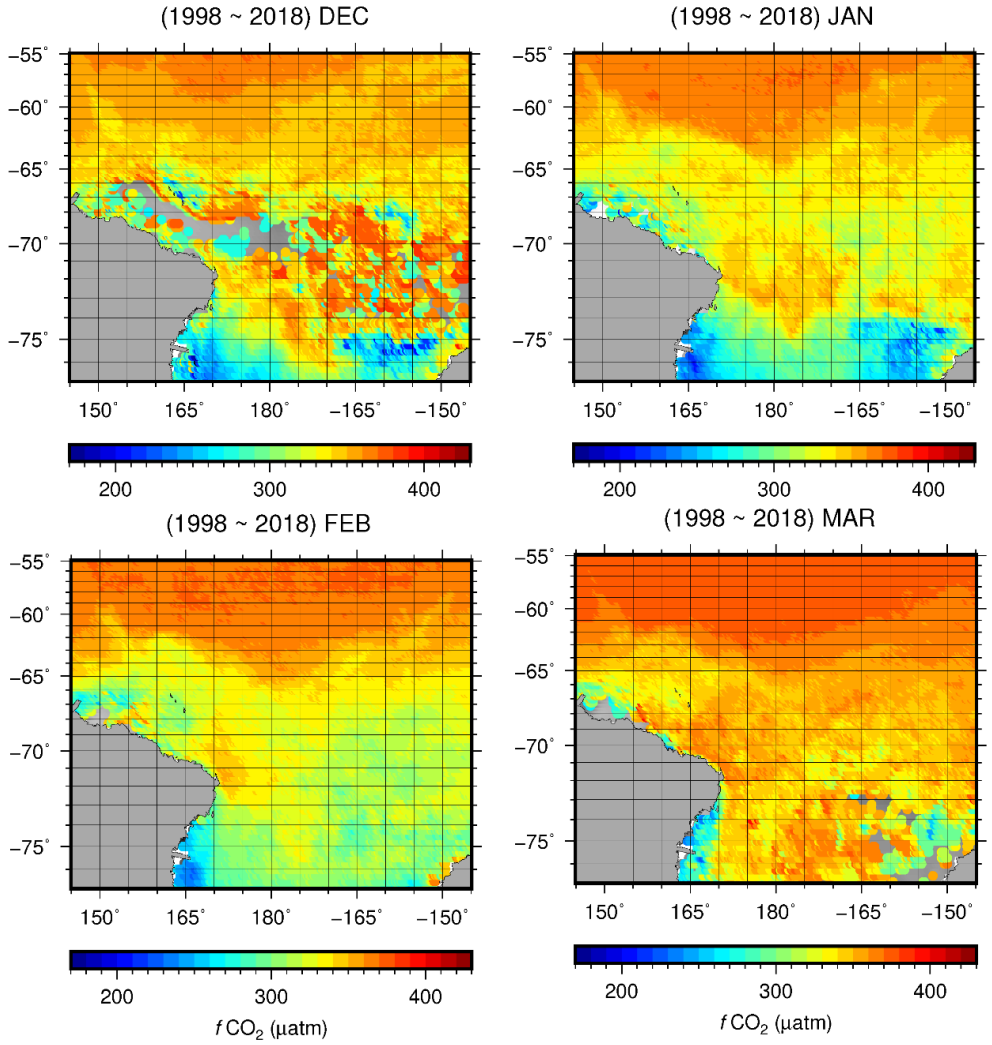

Figure 4. Random forest estimate of 21-year monthly mean fCO_2 in the Ross Sea.

4. 결론 및 토의

본 연구에서는 이산화탄소의 흡수에 중요한 역할을 하는 남극해의 역할의 규명에 있어 중요한 지역인 로스해의 fCO_2의 분포를 기계학습을 통해 성공적으로 재현하였다. 이를 통해, 기존의 관측 부족으로 인한 fCO_2의 흡수와 배출에 대한 로스해 지역의 이해를 증진하는데 활용이 가능할 것으로 판단된다. 또한 나아가, 시간에 따른 이산화탄소의 흡수력의 변동성 평가에도 유용하게 활용이 가능할 것이다. 마지막으로, 기계학습을 이용한 공간적인 분포의 재현을 통해 로스해의 이산화탄소의 흡수에 관한 지역적인 변동성을 확인할 수 있었다. 특히 3월까지 로스해의 동편에서 발생하는 강한 이산화탄소의 흡수패턴은 기존의 보고와 일치하며(DeJong and Dunbar, 2017), 랜덤 포레스트를 통한 재현력의 신뢰성을 더해준다고 보인다. 이외에 3월 로스해의 동편에서도 낮은 fCO_2가 나타나는 것으로 확인되었다. 이는 기존에 보고된 바가 없는 현상으로 향후 로스해의 이산화탄소 흡수력의 평가에 중요한 정보로 활용될 수 있다고 판단한다.

5. 사사

본 연구는 한국해양과학기술원 부설 극지연구소 기관고유사업 "서남극해 온난화에 따른 탄소흡수력 변동 및 생태계 반응 연구 (PE22110)"의 지원을 받아 수행되었습니다. 본 논문의 월별 재현 자료의 고해상도 버전과 데이터 화일은 추후 극지연구소의 데이터관리 시스템(KPDC: https://kpdc.kopri.re.kr/)를 통해 공개할 예정이며, 본 연구의 교신저자에게 이메일 요청 시(keyhongpark@korpi.re.kr) 항시 공유가 가능합니다.

6. 참고문헌

Arrigo KR, van Dijken G, Long M (2008) Coastal Southern Ocean: A strong anthropogenic CO_2 sink. Geophys Res Lett 35(21)

Bakker DC et al. (2016) A multi-decade record of high-quality fCO_2 data in version 3 of the Surface Ocean CO_2 Atlas (SOCAT). Earth Syst Sci Data 8(2): 383-413

Breiman L (2001) Random forests. Mach Learn 45(1): 5-32

Chen S et al. (2019) A machine learning approach to estimate surface ocean pCO_2 from satellite measurements. Remote Sens Environ 228: 203-226

Comiso and Nishio (2008) Trends in the sea ice cover using enhanced and compatible AMSR-E, SSM/I, and SMMR data. J Geophys Res 113: C02S07

DeJong HB, Dunbar RB (2017) Air-sea CO_2 exchange in the Ross Sea, Antarctica. J Geophys Res Oceans 122(10): 8167-8181

Fay AR et al. (2018) Utilizing the Drake Passage Time-series to understand variability and change in subpolar Southern Ocean pCO_2. Biogeosciences 15(12): 3841-3855

Ferry et al. (2012) GLORYS2V1 global ocean reanalysis of the altimetric era (1993-2009) at meso scale. Mercator Ocean Newsletter 44:28-39

Friedlingstein P et al. (2022) Global carbon budget 2021. Earth Syst Sci Data 14(4): 1917-2005

Gregor L, Kok S, Monteiro P (2017) Empirical methods for the estimation of Southern Ocean CO_2: support vector and random forest regression. Biogeosciences 14(23): 5551-5569

Hersbach et al. (2017) Complete ERA5 from 1979: Fifth generation of ECMWF atmospheric reanalysis of the global climate. Copernicus Climate Change Service (C3S) Data Store (CDS)

Huang et al. (2017) Extended reconstructed sea surface temperature version 5 (ERSSTv5): upgrades, validations, and intercomparisons. J Clim 30: 8179-8205

Landshuster et al. (2014) Recent variability of the global ocean carbon sink. Global Biogeochem Cycles 28(9): 927-949

Ma L, Fan S (2017) CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. BMC Bioinform 18(1): 1-18

Maritorena, Siegel (2005) Consistent merging of satellite ocean color data sets using a bio-optical model. Remote Sens Environ 94: 429-440

Nakaoka et al. (2013) Estimating temporal and spatial variation of ocean surface pCO_2 in the North Pacific using a self-organizing map neural network technique. Biogeosciences 10(9): 6093-6106

Reynolds et al. (2008) NOAA optimum interpolation 1/4 degree daily sea surface temperature (OISST) analysis, version 2. NOAA National Centers for Environmental Information. doi:10.7289/V5SQ8XB5

Sutton AJ, Williams NL, Tilbrook B (2021) Constraining Southern Ocean CO_2 flux uncertainty using uncrewed surface vehicle observations. Geophys Res Lett 48(3): e2020GL091748

Williams NL et al. (2017) Calculating surface ocean pCO_2 from biogeochemical Argo floats equipped with pH: An uncertainty analysis. Global Biogeochem Cycles 31(3): 591-604

Zeng et al. (2014) A global surface ocean fCO_2 climatology based on a feed-forward neural network. J Atmos Ocean Technol 31(8): 1838-1849

7. 데이터셋에 대한 메타데이터

Sort	Field	Subcategory#1	Subcategory#2
Essential	*Title	Monthly fCO_2 from 1998 to 2018 in the Ross Sea	
	*DOI name	10.22761/DATA2022.4.3.003	
	*Category	Oceans	
	Abstract	Reproduced fCO_2 distribution of sea surface water using machine learning technique	
	*Temporal Coverage	From 1998, January to 2018, December	
	*Spatial Coverage	Latitude: 55.00°S – 72.00°S Longitude: 145.00°E – 145.00°W	
	*Personnel	Name	Keyhong Park
		Affiliation	Korea Polar Research Institute (KOPRI)
		E-mail	keyhongpark@kopri.re.kr
	*CC License	CC BY-NC	
Optional	*Project	서남극해 온난화에 따른 탄소흡수력 변동 및 생태계 반응 연구	PE22110