A Machine Learning Method for Estimating Atmospheric Trace Gas Concentration Baselines

Kirstin Gerrand[1, 2], Elena Fillola[1, 3], Alistair J. Manning[4, 1], Jgor Arduini[5], Paul B. Krummel[6], Chris R. Lunder[7], Jens Mühle[8], Simon O'Doherty[1], Sunyoung Park[9], Ronald G. Prinn[10], Stefan Reimann[11], Dickon Young[1], and Matthew Rigby[1]

[1]Atmospheric Chemistry Research Group, School of Chemistry, University of Bristol, Bristol, UK
[2]New Zealand Institute for Earth Science, Wellington, New Zealand
[3]School of Engineering Mathematics, University of Bristol, Bristol, UK
[4]Hadley Centre, Met Office, Exeter, UK
[5]Department of Pure and Applied Sciences, University of Urbino, Urbino, 61029, Italy
[6]CSIRO Environment, Aspendale, VIC, Australia
[7]NILU, Kjeller, Norway
[8]Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA
[9]Kyungpook Institute of Oceanography, Kyungpook National University, Daegu, Republic of Korea
[10]Center for Sustainability Science and Strategy, Massachusetts Institute of Technology, Cambridge, MA, USA
[11]Laboratory for Air Pollution/Environmental Technology, Empa, Dübendorf, Switzerland

Supplementary Material

Here, we show additional figures and tables that elaborate on model choices and performance.

An example comparison of ECMWF and UK Met Office wind speed and direction at Mace Head, Ireland is shown in Figure S1, indicating that differences between the meteorology used to drive the NAME model (UK Met Office analysis) and that used in training our algorithm (ECMWF ERA5) are small and unlikely to contribute substantially to algorithm performance issues.

A summary of the meteorological parameters used as features in the machine learning model are provided in Table S1.

Metrics derived from the Random Forest (RF) model are provided in Table S2, based on hyperparameter sets in Tables S3a and S3b, and the the RF confusion matrix and heatmap is shown in Figs. S2 and S3, respectively.

Hyperparameter sets for the final MLP models are shown in Tables S4a and S4b. The feature importance for the MLP models are shown in Table S5. Plots indicating MLP model performance for a range of species and sites are shown in Section 4.3.

1 Intercomparison of ECMWF and Met Office UM Meteorological Fields

Figure S1. Subplots comparing ECMWF and Met Office UM 10m winds across a six-month sample period (Jan-June 2015) at Mace Head, Ireland. Top figure compares wind directions, and bottom, wind speeds. Legend contains average values for additional quantitative comparison.

2 Model Inputs

Table S1. Input variable categories used in the ML models. Each variable is extracted from 17 horizontally distributed points around the site; the data collection site itself and 16 from surrounding nearest grid points based on two 3x3 grids spanning $\pm$ 5 $^{\circ}$ and $\pm$ 10 $^{\circ}$ latitude and longitude.

Category	*Units*
Boundary layer height	m
Surface Pressure	Pa
10m u-wind	m s^{-1}
10m u-wind (T-6)	m s^{-1}
500hPa u-wind	m s^{-1}
500hPa u-wind (T-6)	m s^{-1}
850hPa u-wind	m s^{-1}
850hPa u-wind (T-6)	m s^{-1}
10m v-wind	m s^{-1}
10m v-wind (T-6)	m s^{-1}
500hPa v-wind	m s^{-1}
500hPa v-wind (T-6)	m s^{-1}
850hPa v-wind	m s^{-1}
850hPa v-wind (T-6)	m s^{-1}
Time of Day	
Day of Year	

3 Random Forest Summary

3.1 Final Random Forest Model Outcomes

Table S2. A tabular summary of the final RF model outcomes, showing precision, recall and F1 score values for each of the AGAGE sites. Scores can range from 0 to 1, with higher scores indicating higher performance.

Station Designation	*Precision*	*Recall*	*F1 score*
CGO	0.953	0.818	0.880
CMN	0.957	0.588	0.728
GSN	0.925	0.608	0.734
JFJ	0.941	0.674	0.785
MHD	0.917	0.672	0.776
RPB	0.884	0.521	0.655
SMO	0.896	0.404	0.557
THD	0.915	0.437	0.591
ZEP	0.948	0.718	0.817

3.2 Final Hyperparameter Sets

Note that hyperparameters not listed are as default in the documentation.

Table S3. Hyperparameter sets used in the final versions of the RF models for AGAGE sites.

(a) Four AGAGE sites: Kennaook/Cape Grim (CGO), Monte Cimone (CMN), Gosan (GSN), Jungfraujoch (JFJ).

Parameter	*CGO*	*CMN*	*GSN*	*JFJ*
`random_state`	42	42	42	42
`n_estimators`	100	100	200	100
`max_depth`	5	3	5	5
`criterion`	'entropy'	'entropy'	'entropy'	'entropy'
`bootstrap`	False	False	False	False
`max_features`	'sqrt'	'sqrt'	None	'sqrt'

(b) Five AGAGE sites: Mace Head (MHD), Ragged Point (RPB), Cape Matatula (SMO), Trinidad Head (THD), Zeppelin (ZEP).

Parameter	*MHD*	*RPB*	*SMO*	*THD*	*ZEP*
`random_state`	42	42	42	42	42
`n_estimators`	200	100	50	100	200
`max_depth`	8	5	5	5	5
`criterion`	'entropy'	'entropy'	'log_loss'	'gini'	'entropy'
`bootstrap`	False	False	False	True	False
`min_samples_split`	5	5	3	2	2
`max_features`	'sqrt'	'log2'	'sqrt'	'sqrt'	'sqrt'

3.3 Random Forest Confusion Matrix Map

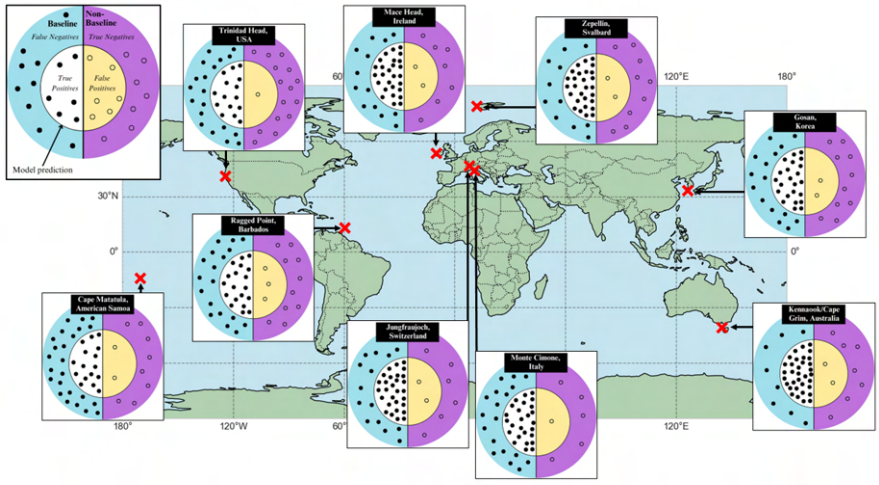


Figure S2. A map showing the locations of the nine AGAGE sites, with confusion matrix-derived plots at each location. Each confusion matrix was normalised, to remove visual differences due to testing set sizes; each point represents approximately 2% of the total test set (rounding means that the total number of points on each plot range from 49 to 51). The left half of each circle represents true baseline points, and the right half true non-baseline points. The inner circle shows the random forest model prediction of baseline points. The key in the top left indicates where true or false positives and negatives lie, as explained in the main text.

3.4 Random Forest MAPE Heatmap

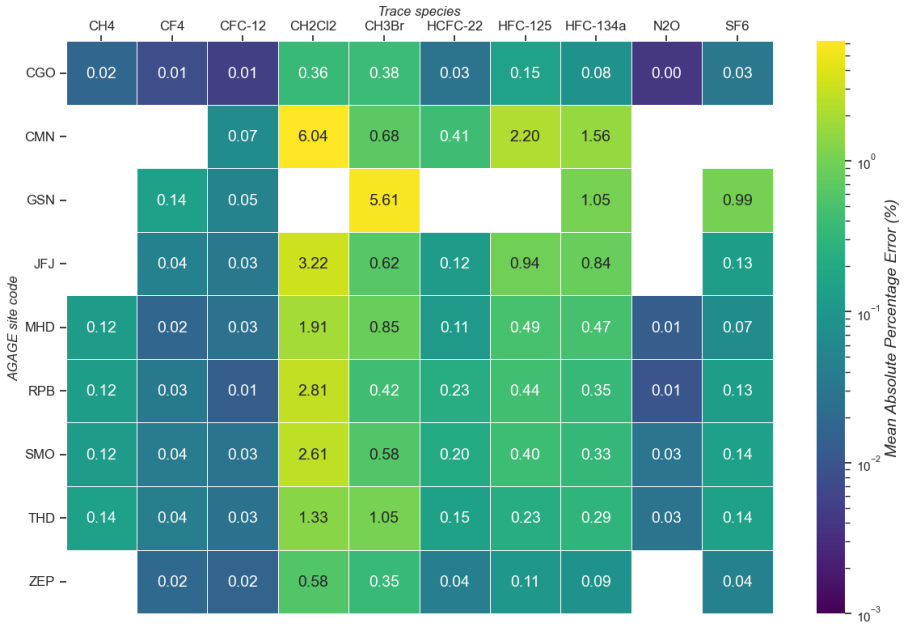


Figure S3. Mean Absolute Percentage Errors (MAPE) for the random forest model for selected AGAGE species across the nine monitoring sites. White denotes species that were not available at a particular site.

20 **4 MLP Summary**

4.1 Final Hyperparameter Sets

Note that hyperparameters not listed are as default in the documentation.

Table S4. Hyperparameter sets used in the final versions of the MLP models for AGAGE sites.

(a) Four AGAGE sites: Kennaook/Cape Grim (CGO), Monte Cimone (CMN), Gosan (GSN), Jungfraujoch (JFJ).

Parameter	*CGO*	*CMN*	*GSN*	*JFJ*
`random_state`	42	42	42	42
`max_iter`	1000	1000	1000	50
`hidden_layer_sizes`	(50,)	(100,)	(100,)	(25,)
`shuffle`	False	False	False	False
`alpha`	0.05	0.5	0.0001	0.1
`learning_rate`	'constant'	'constant'	'constant'	'invscaling'
`batch_size`	100	250	100	100
`early_stopping`	True	False	False	True
`learning_rate_init`	0.001	0.01	0.01	0.01
`epsilon`	1e-8	1e-8	1e-8	2e-10
`beta_2`	0.9	0.9	0.9	0.8
`tol`	1e-4	1e-4	1e-4	1e-3

(b) Five AGAGE sites: Mace Head (MHD), Ragged Point (RPB), Cape Matatula (SMO), Trinidad Head (THD), Zeppelin (ZEP).

Parameter	*MHD*	*RPB*	*SMO*	*THD*	*ZEP*
`random_state`	42	1	42	42	42
`max_iter`	1000	500	1000	1000	1000
`hidden_layer_sizes`	(200,)	(200,)	(100,)	(100,)	(200,150,)
`shuffle`	False	False	False	False	False
`alpha`	0.05	0.0001	0.001	0.05	0.05
`learning_rate`	'constant'	'constant'	'constant'	'constant'	'constant'
`batch_size`	100	100	100	100	100
`early_stopping`	True	True	True	True	True
`learning_rate_init`	0.01	0.001	0.001	0.001	0.01
`beta_2`	0.9	0.9	0.9	0.9	0.9

4.2 Feature Importance

Table S5. A tabular summary of the top three most influential feature groups when training the final MLP model for each site. u- and v- wind encompass all heights, locations and times. 'PBLH' represents planetary boundary layer height.

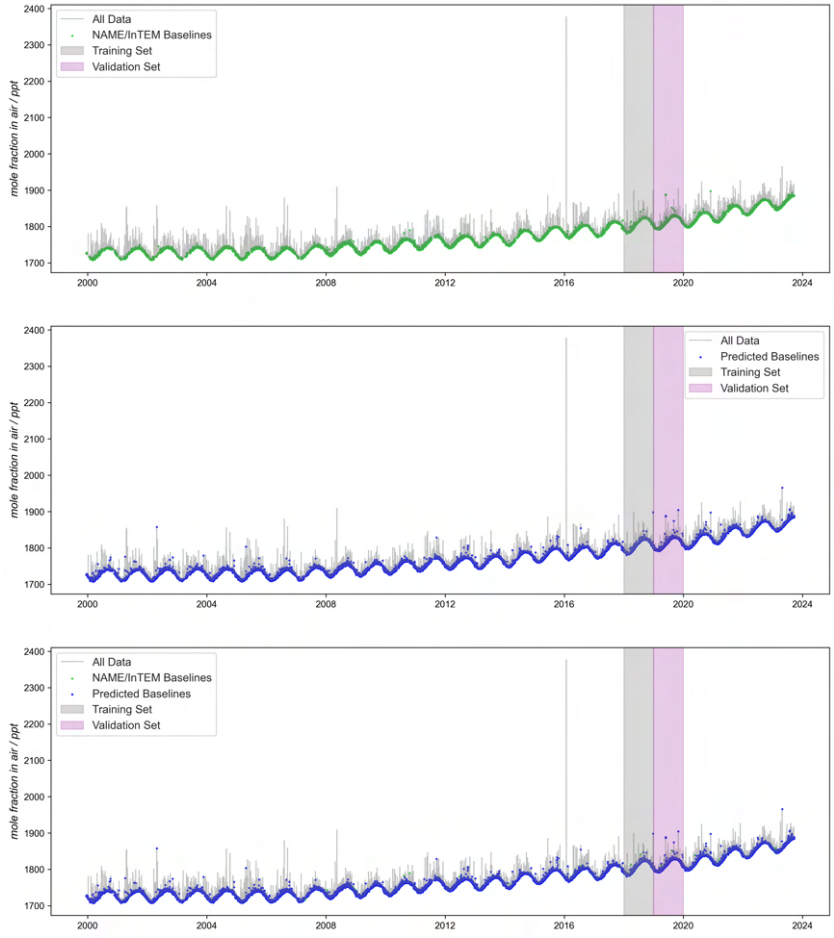
Station Designation	*Most Importance*	*Second*	*Third*
CGO	u-wind	v-wind	PBLH
CMN	PBLH	u-wind	Surface pressure
GSN	PBLH	u-wind	Surface pressure
JFJ	v-wind	u-wind	Surface Pressure
MHD	u-wind	v-wind	PBLH
RPB	u-wind	v-wind	PBLH
SMO	u-wind	w-wind	PBLH
THD	PBLH	v-wind	u-wind
ZEP	v-wind	u-wind	PBLH

4.3 MLP Plots

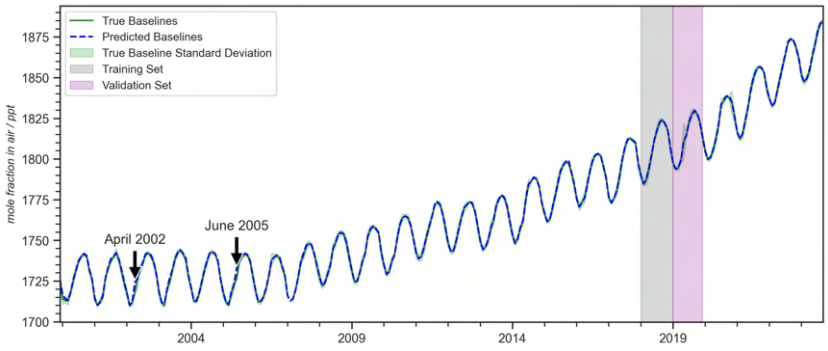
4.4 Kennaook/Cape Grim, Australia

4.4.1 CH_4

Mole fraction time series

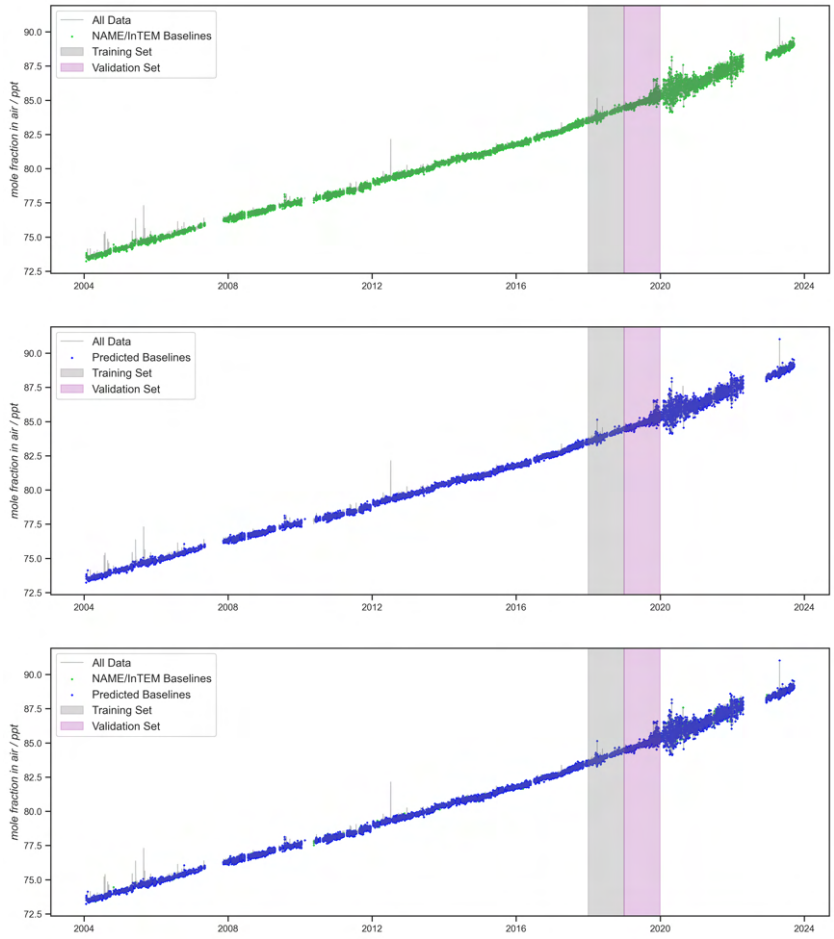


Monthly means

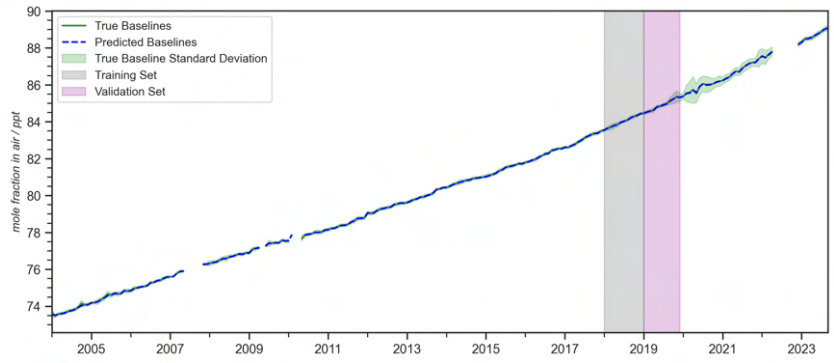


4.4.2 CF_4

Mole fraction time series

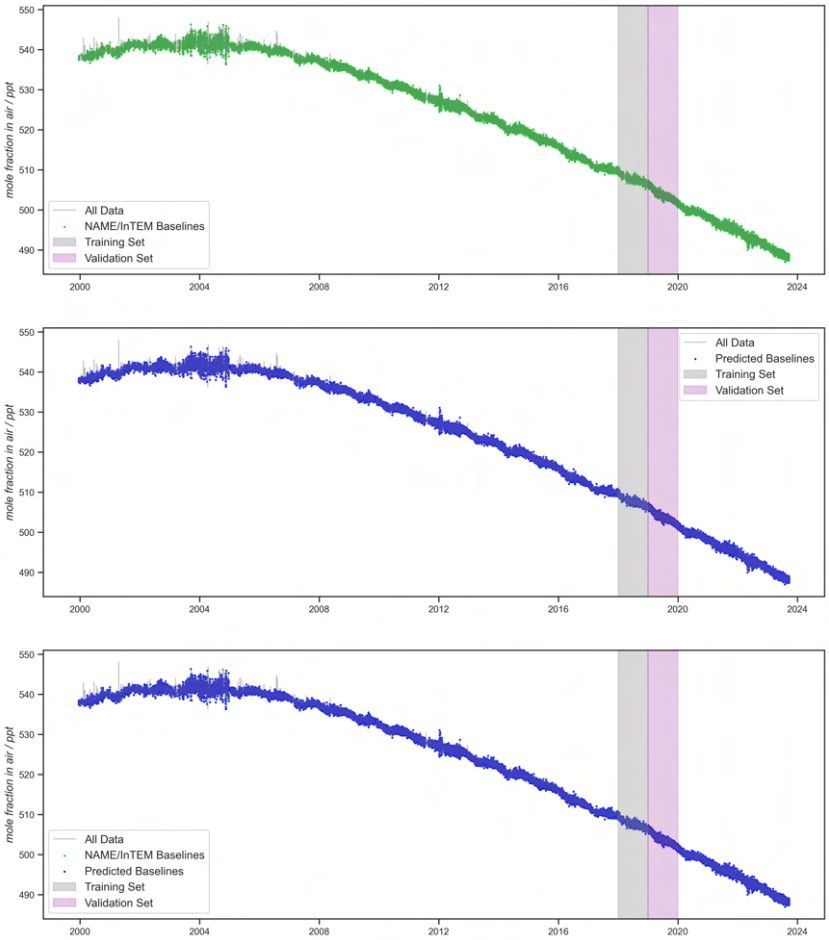


Monthly means

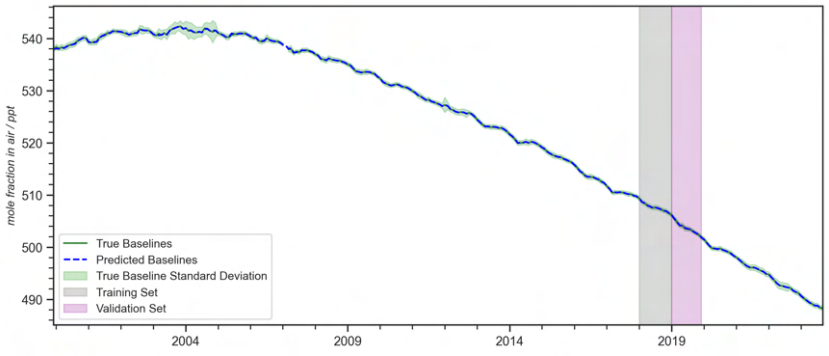


4.4.3 CFC-12

Mole fraction time series

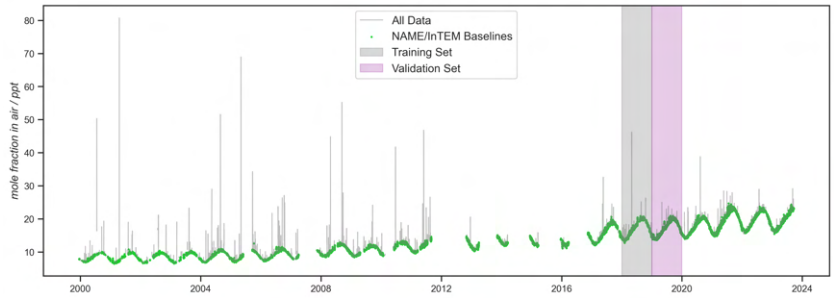


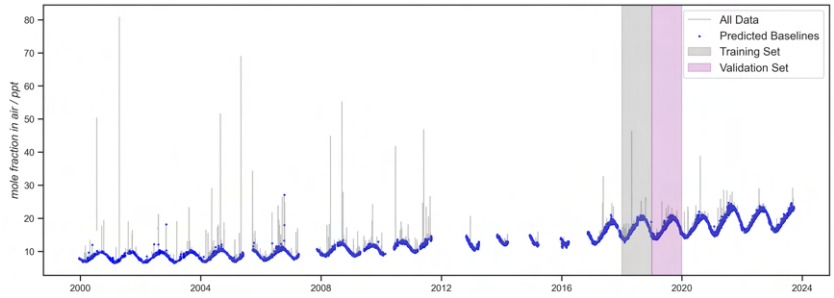
Monthly means

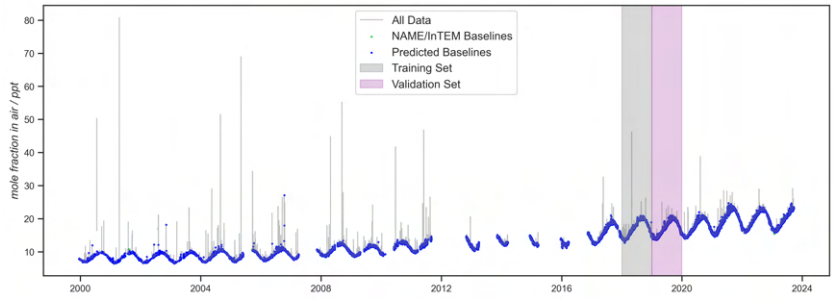


4.4.4 CH_2Cl_2

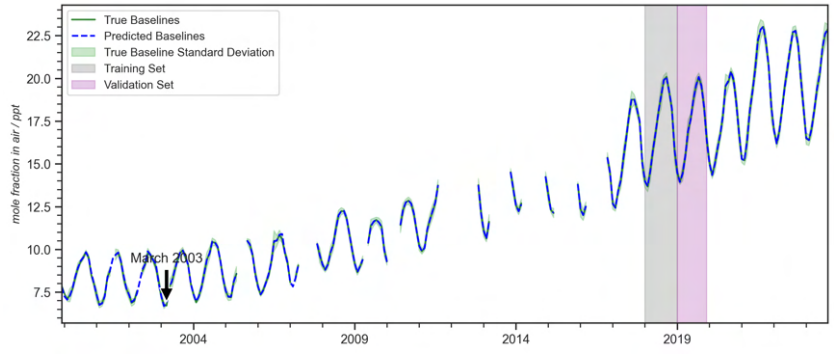
Mole fraction time series





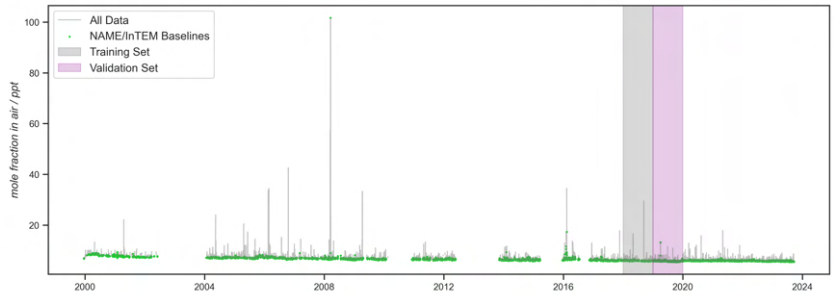


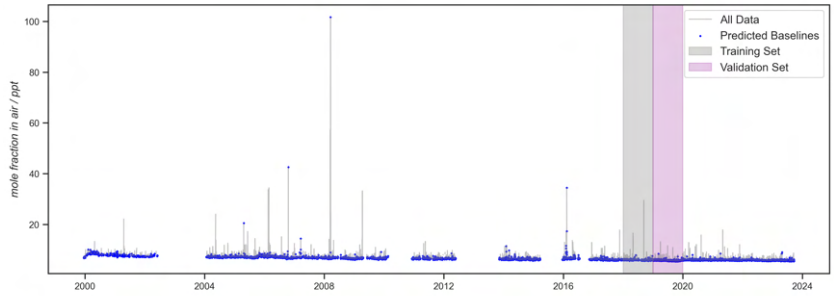
Monthly means

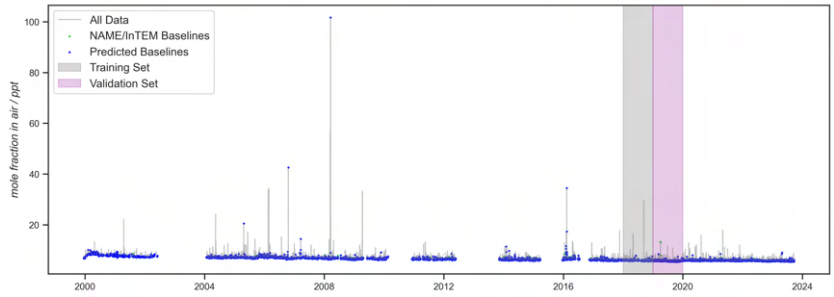


4.4.5 CH_3Br

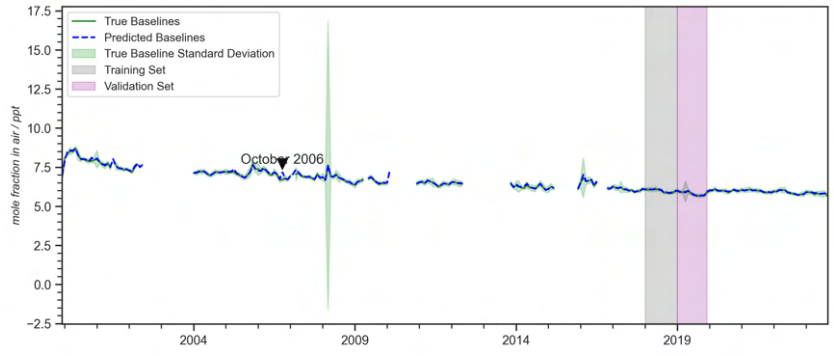
Mole fraction time series





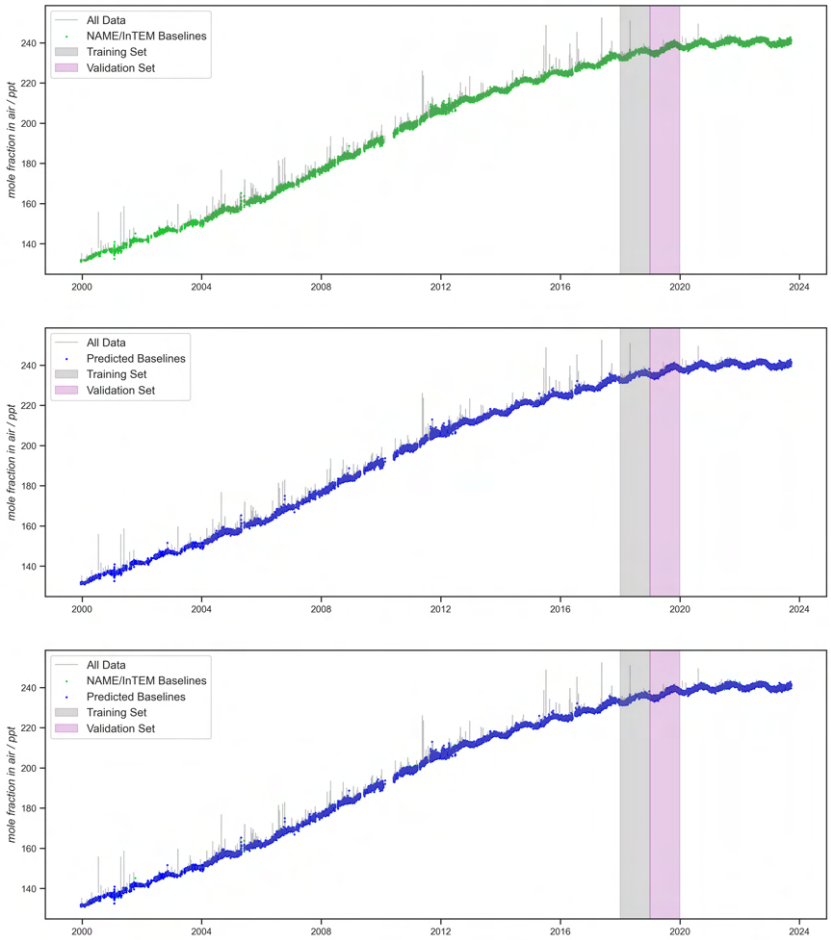


Monthly means

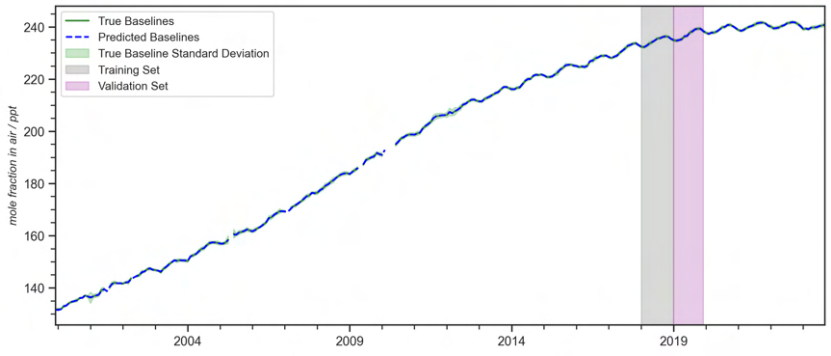


4.4.6 HCFC-22

Mole fraction time series

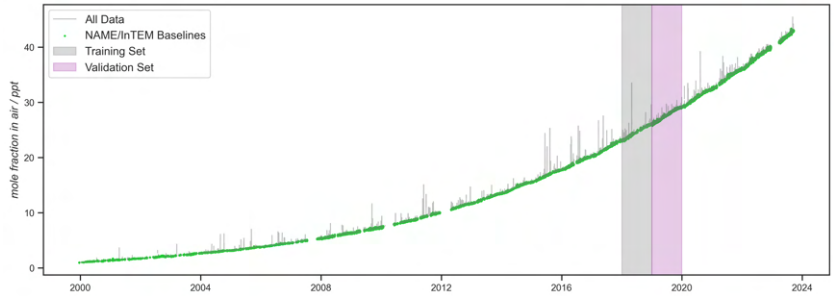


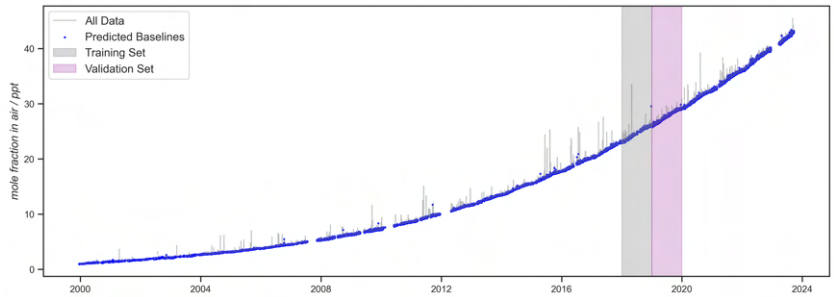
Monthly means

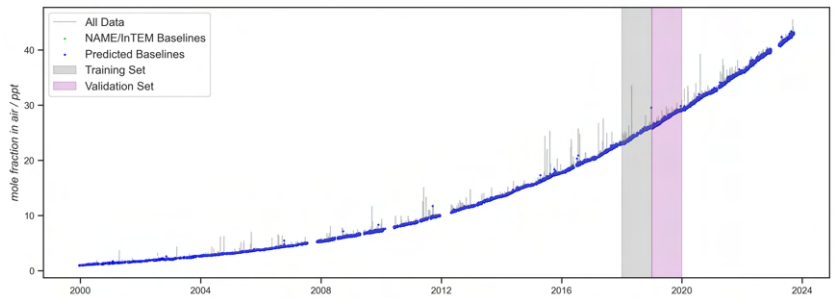


4.4.7 HFC-125

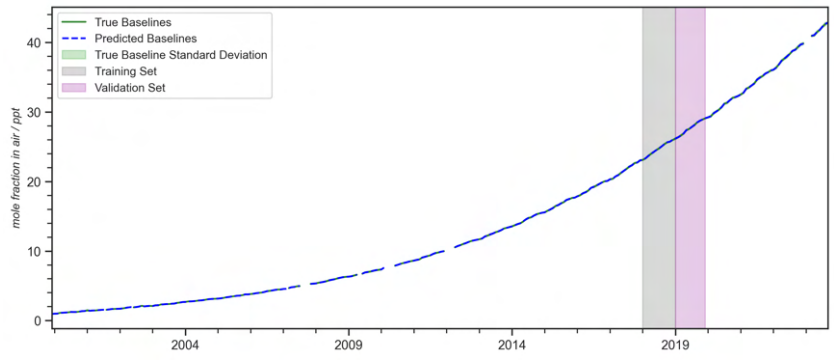
Mole fraction time series





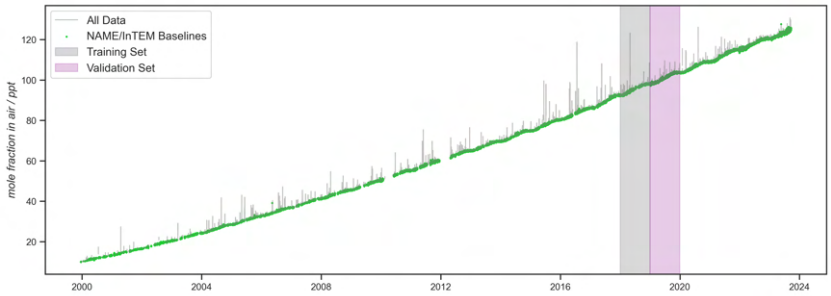


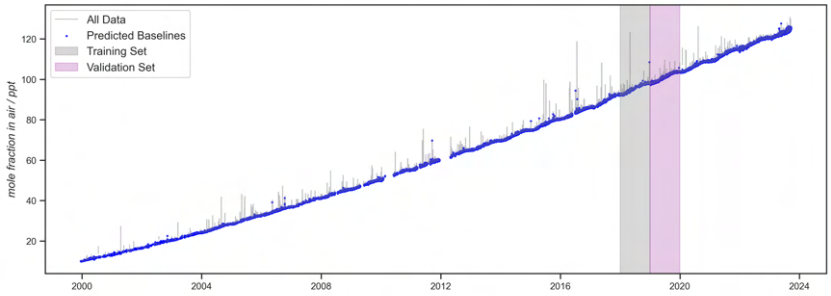
Monthly means

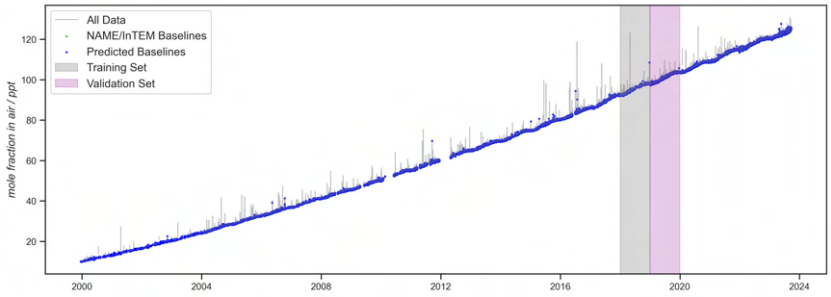


4.4.8 HFC-134a

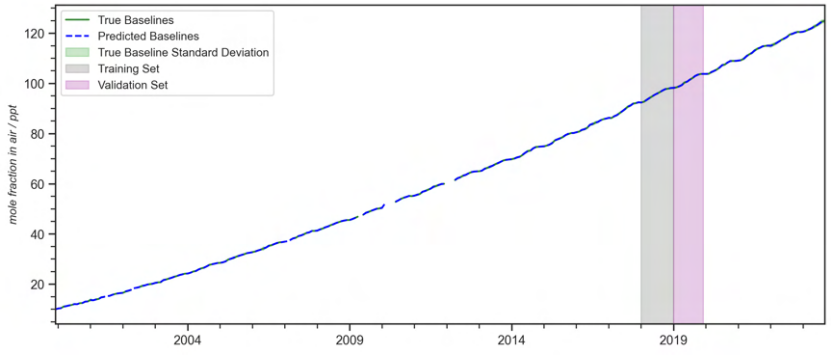
Mole fraction time series





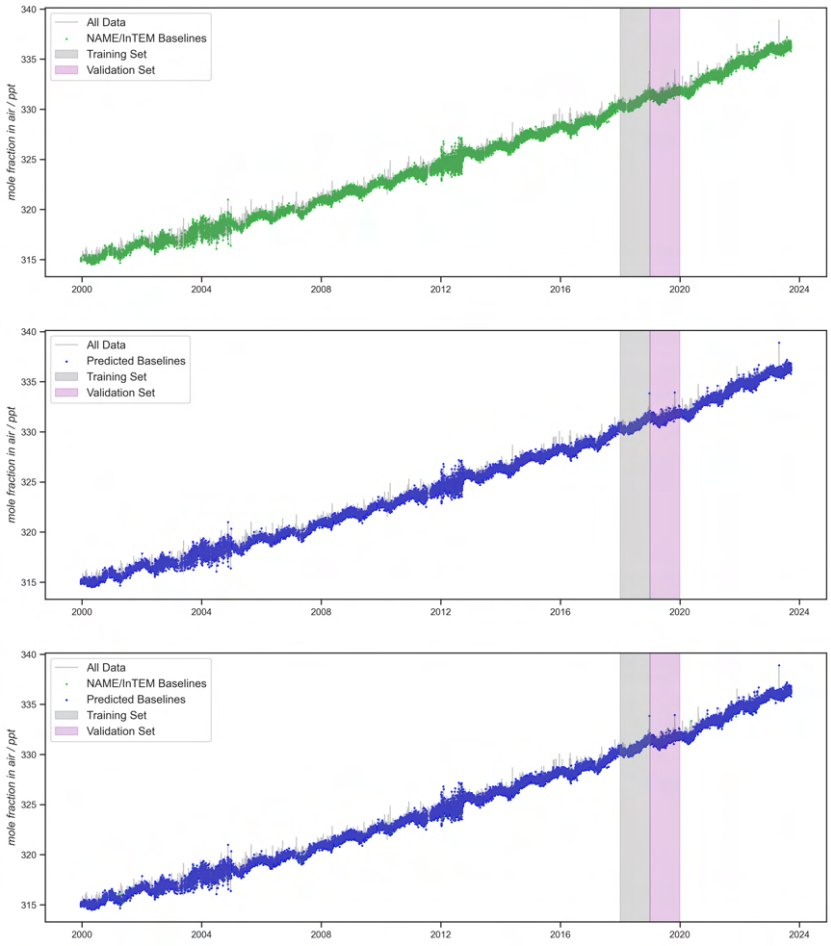


Monthly means

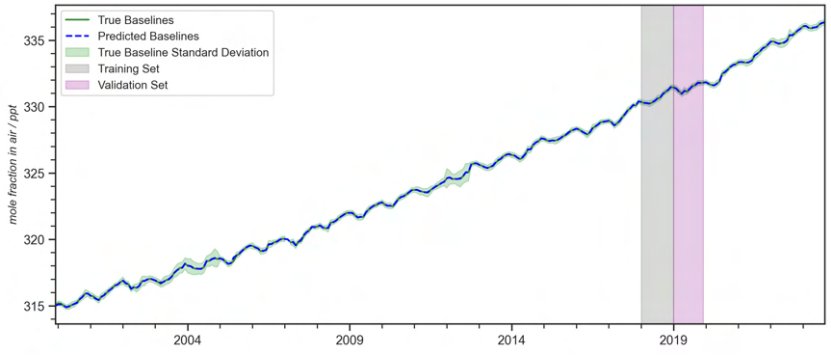


4.4.9 N_2O

Mole fraction time series

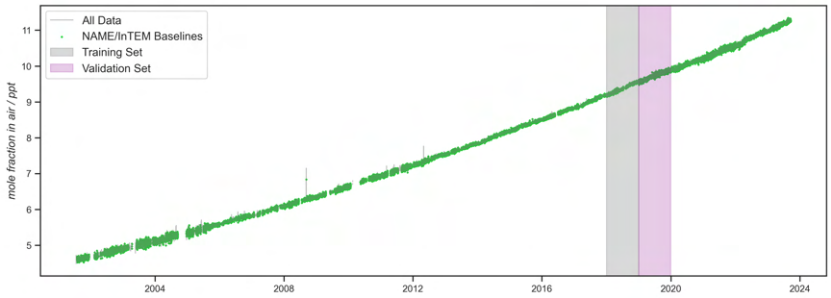


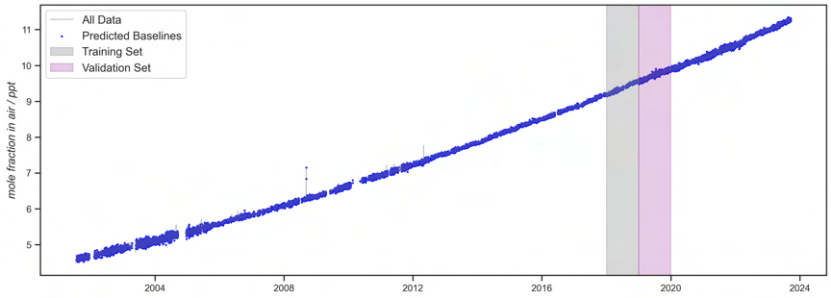
Monthly means

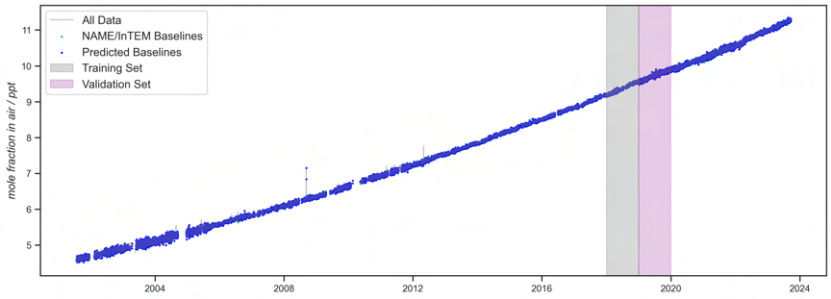


4.4.10 SF_6

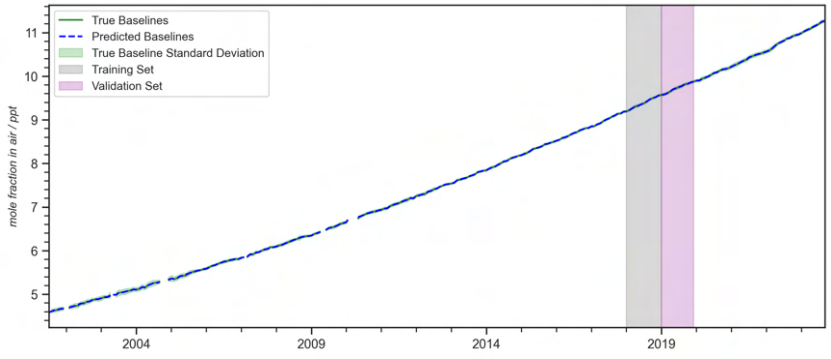
Mole fraction time series







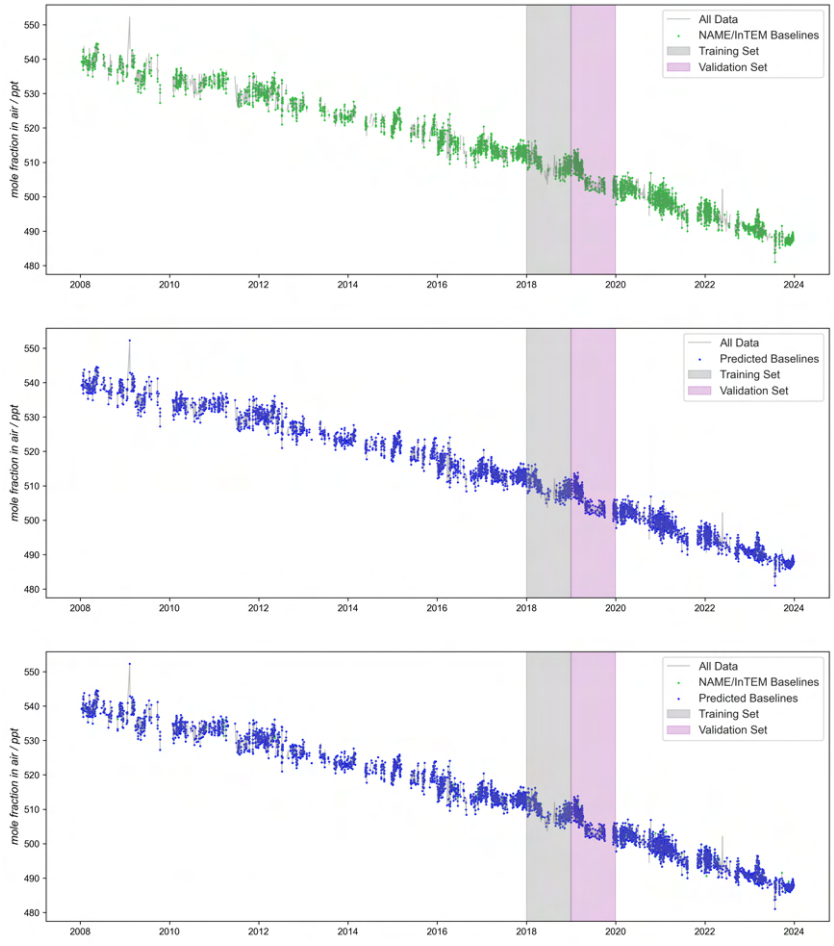
Monthly means



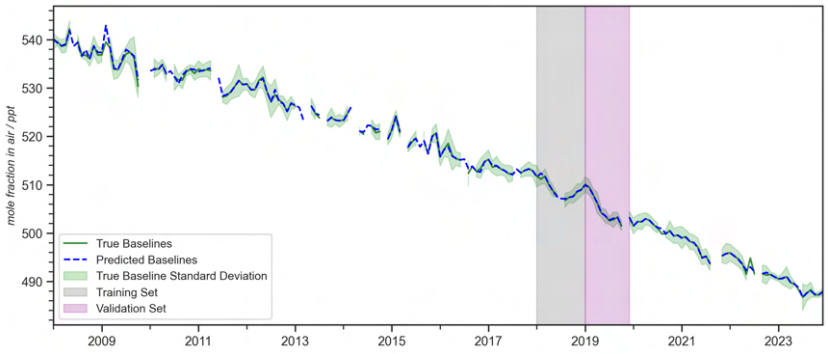
4.5 Monte Cimone, Italy

4.5.1 CFC-12

Mole fraction time series

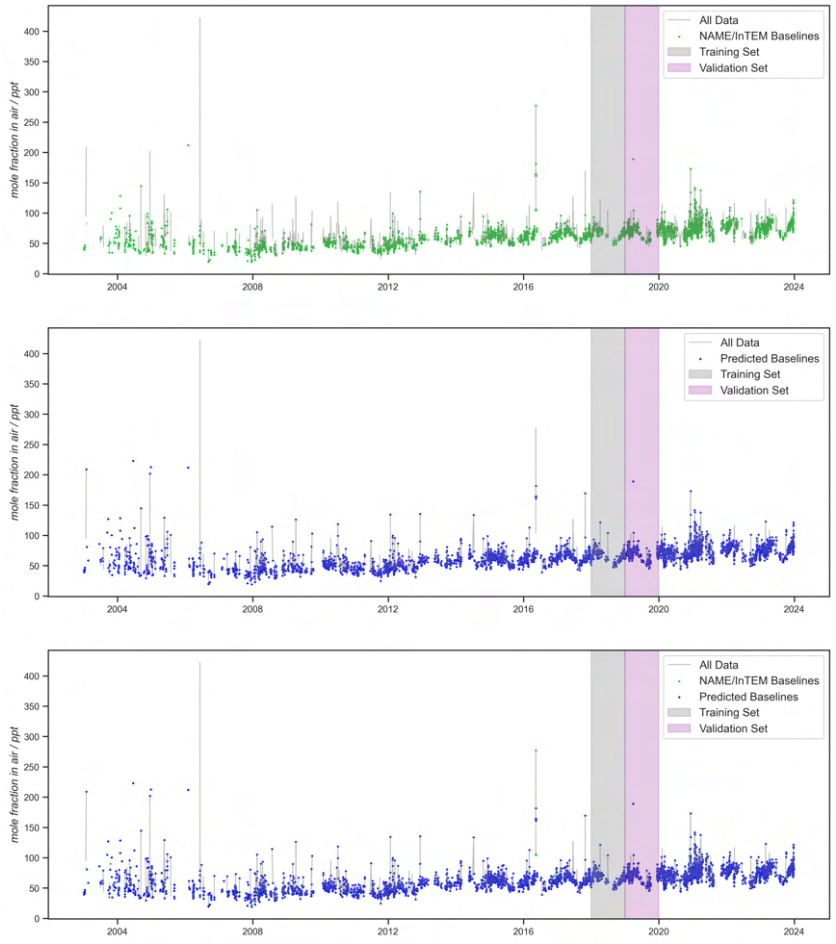


Monthly means

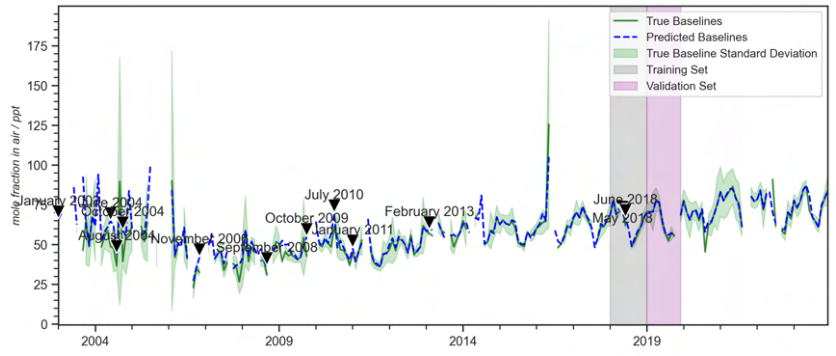


4.5.2 CH_2Cl_2

Mole fraction time series

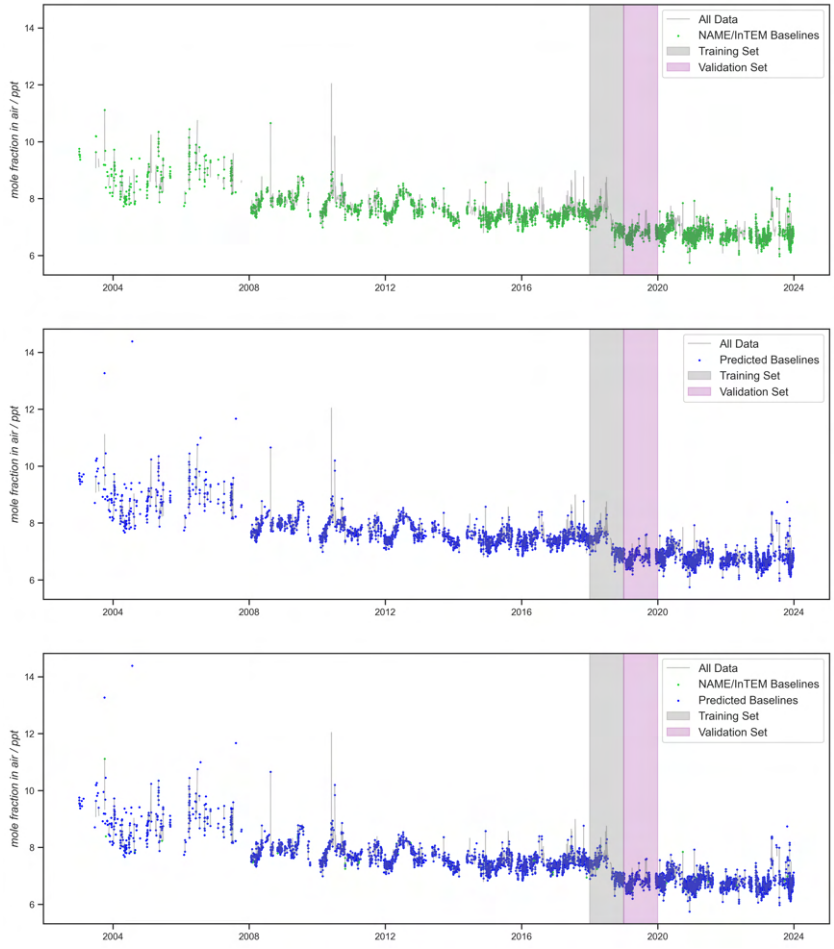


Monthly means

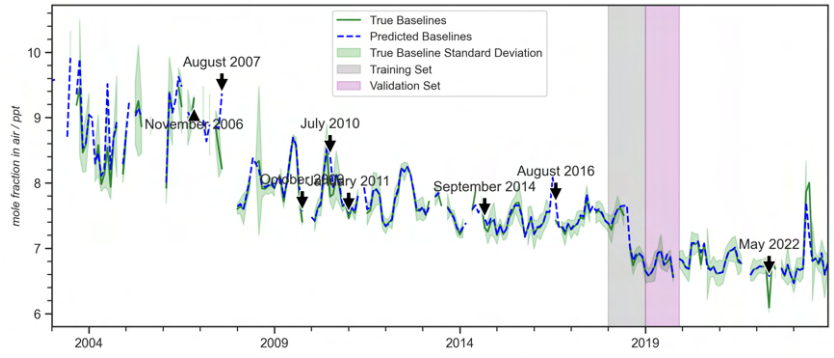


4.5.3 CH_3Br

Mole fraction time series

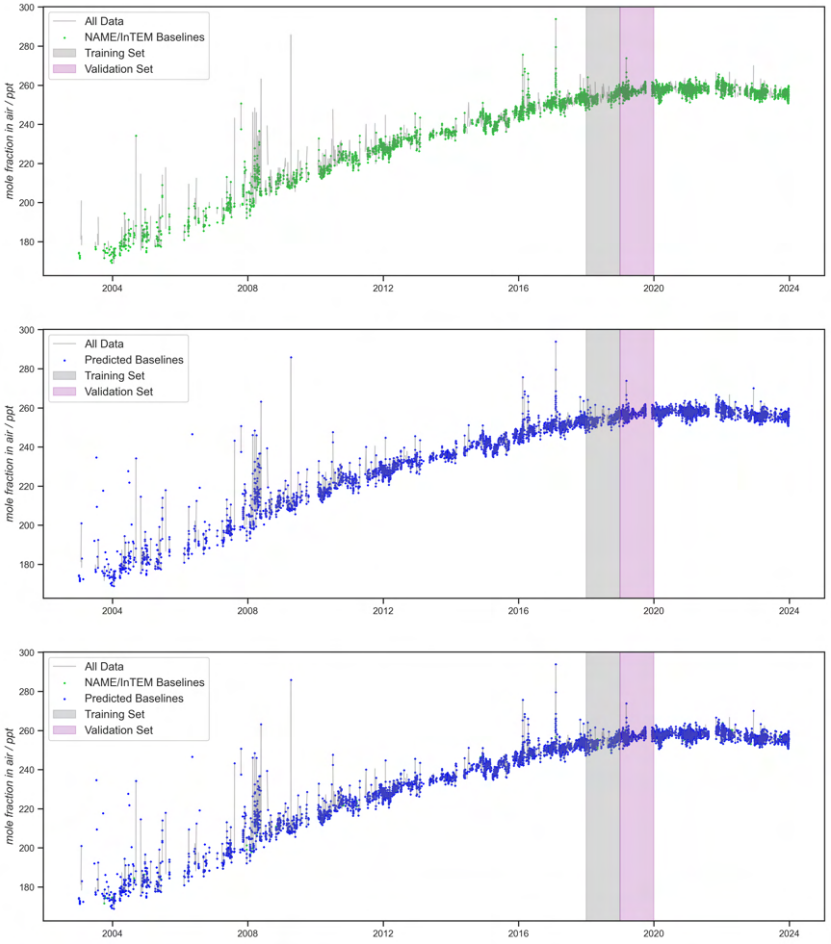


Monthly means

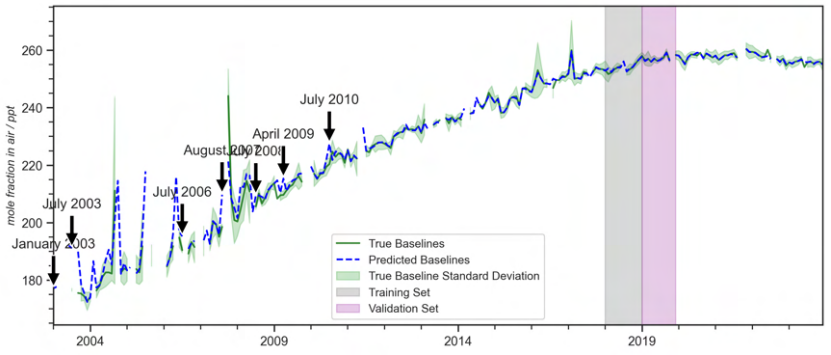


4.5.4 HCFC-22

Mole fraction time series

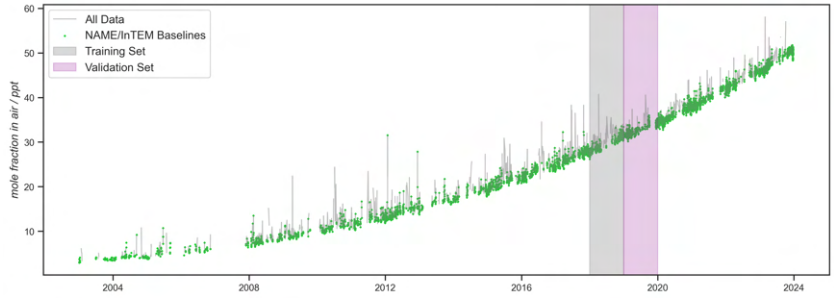


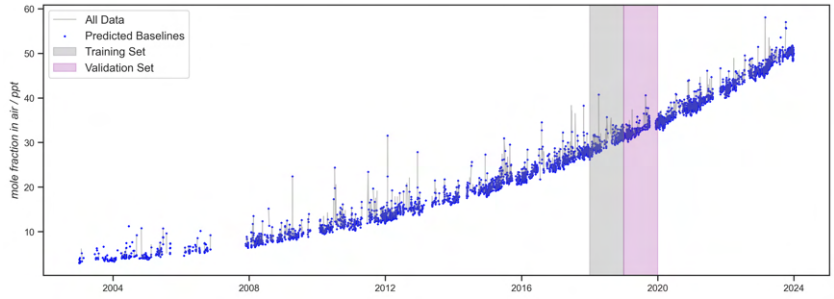
Monthly means

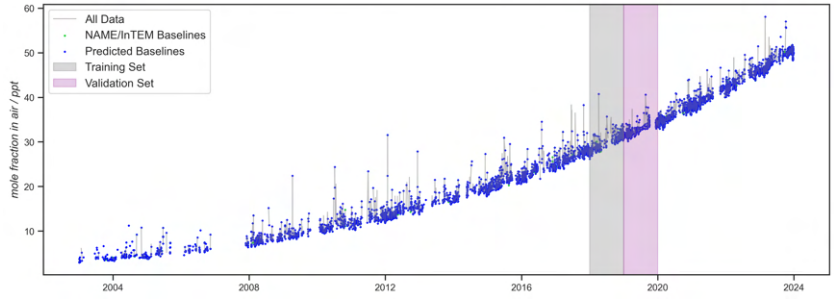


4.5.5 HFC-125

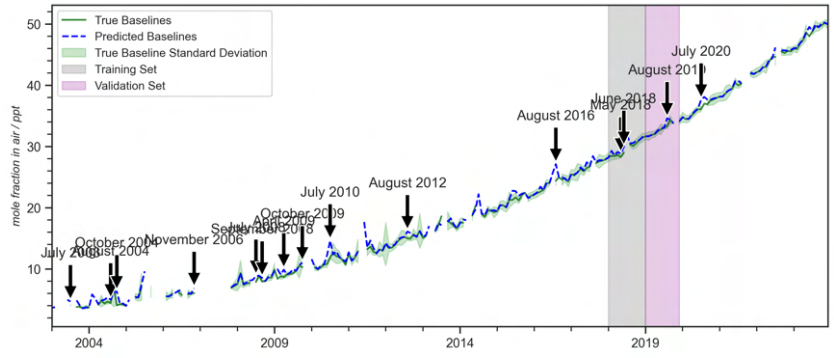
Mole fraction time series





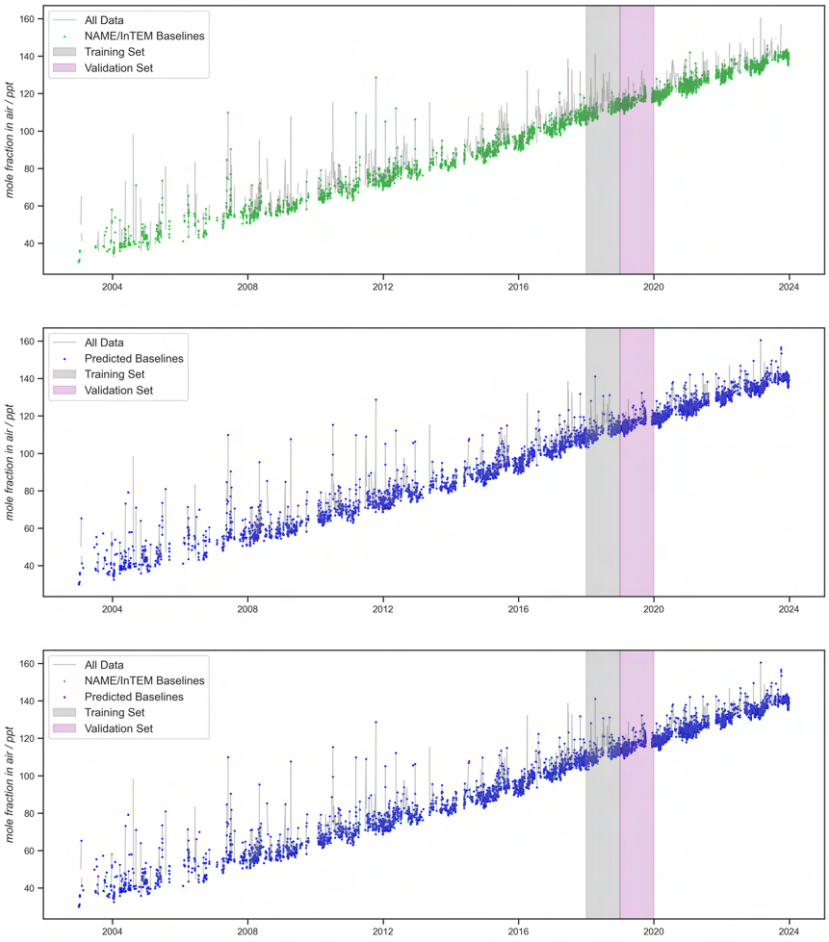


Monthly means

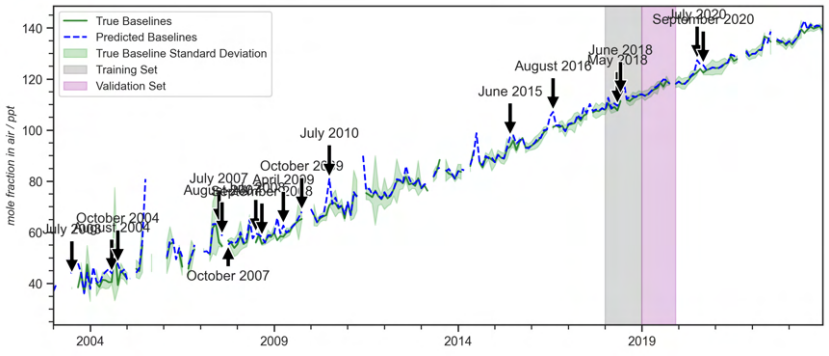


4.5.6 HFC-134a

Mole fraction time series



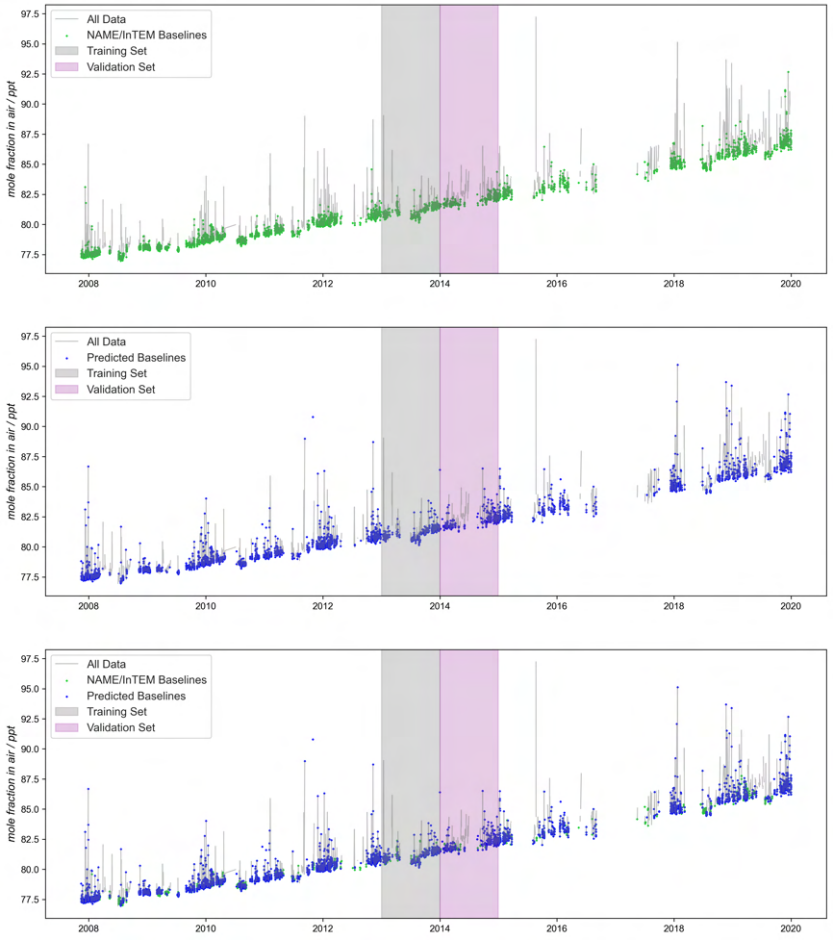
Monthly means



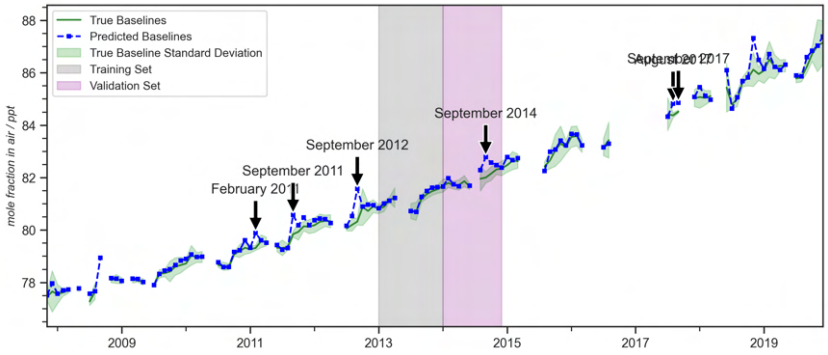
4.6 Gosan, South Korea

4.6.1 CF_4

Mole fraction time series

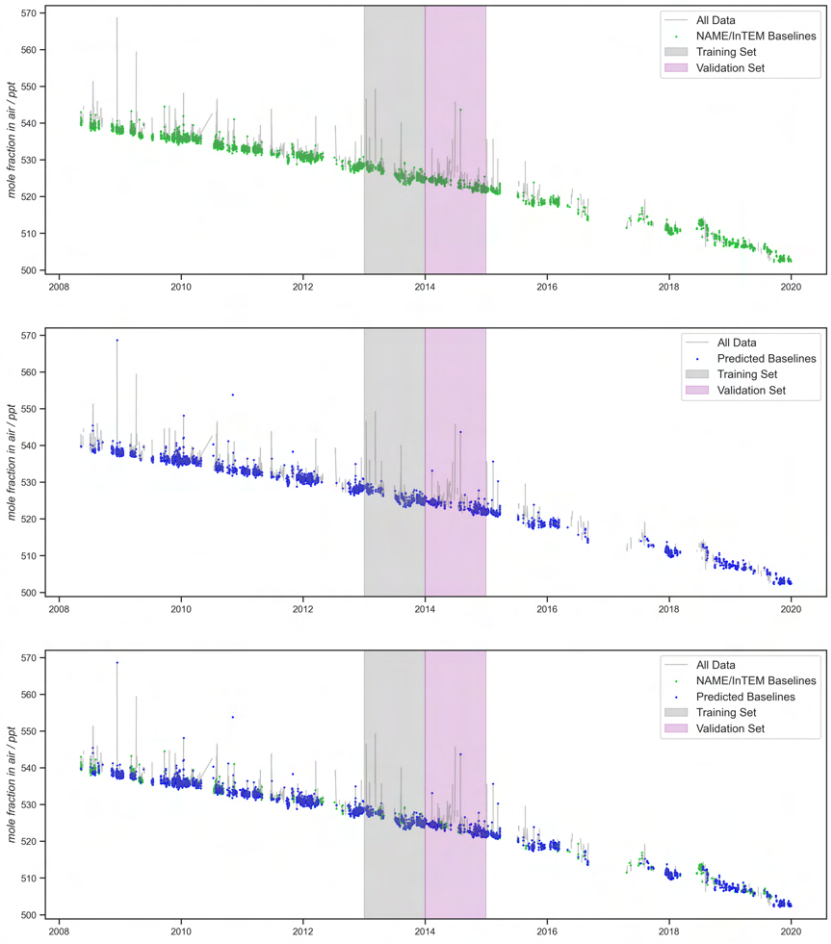


Monthly means

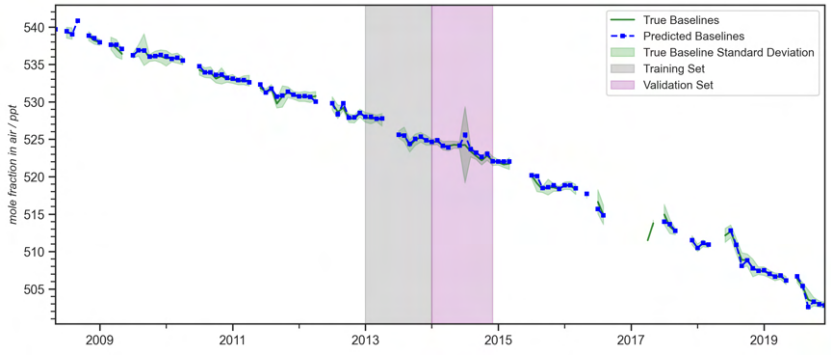


4.6.2 CFC-12

Mole fraction time series

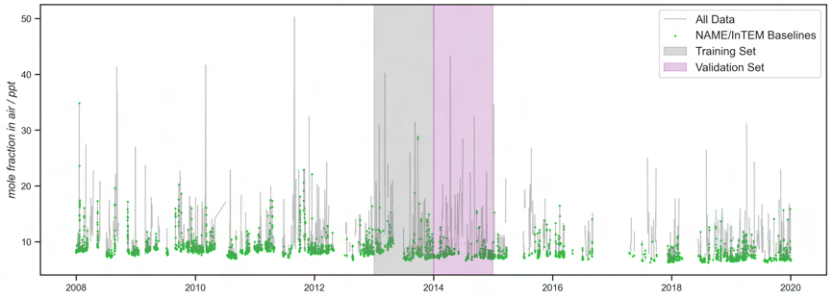


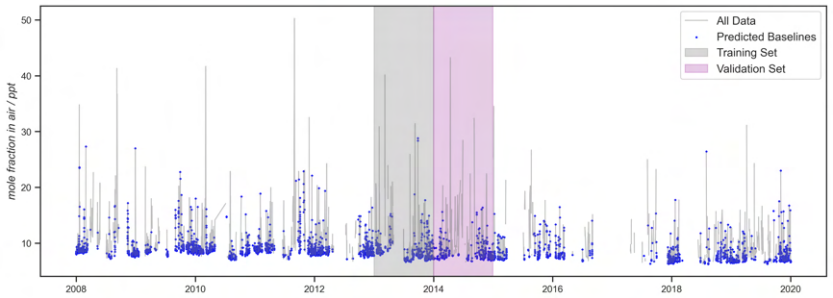
Monthly means

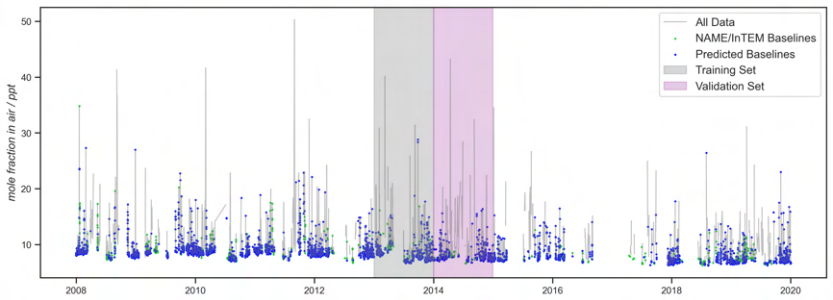


4.6.3 CH_3Br

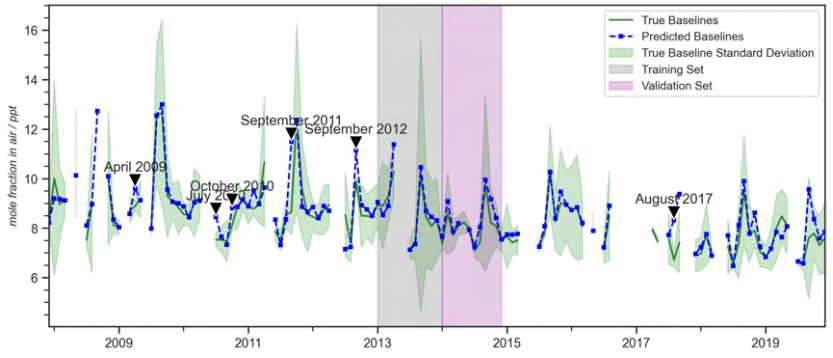
Mole fraction time series





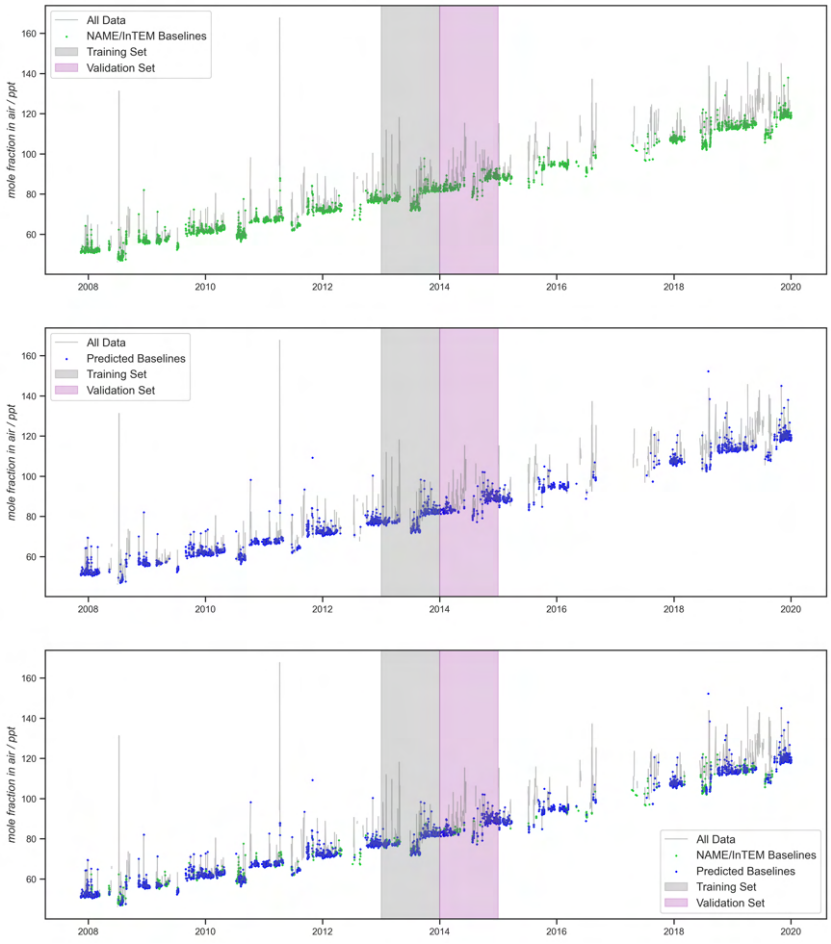


Monthly means

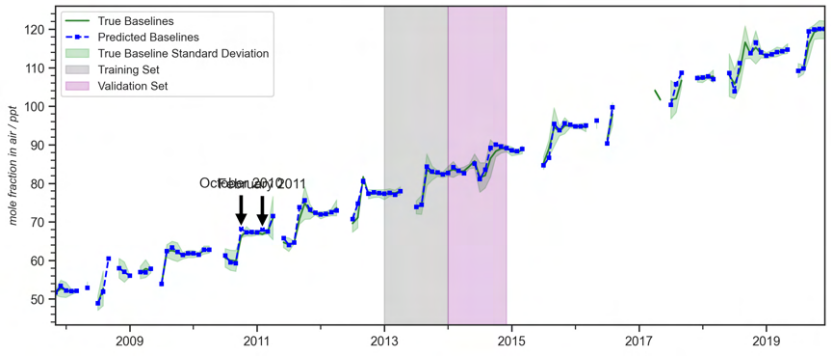


4.6.4 HFC-134a

Mole fraction time series

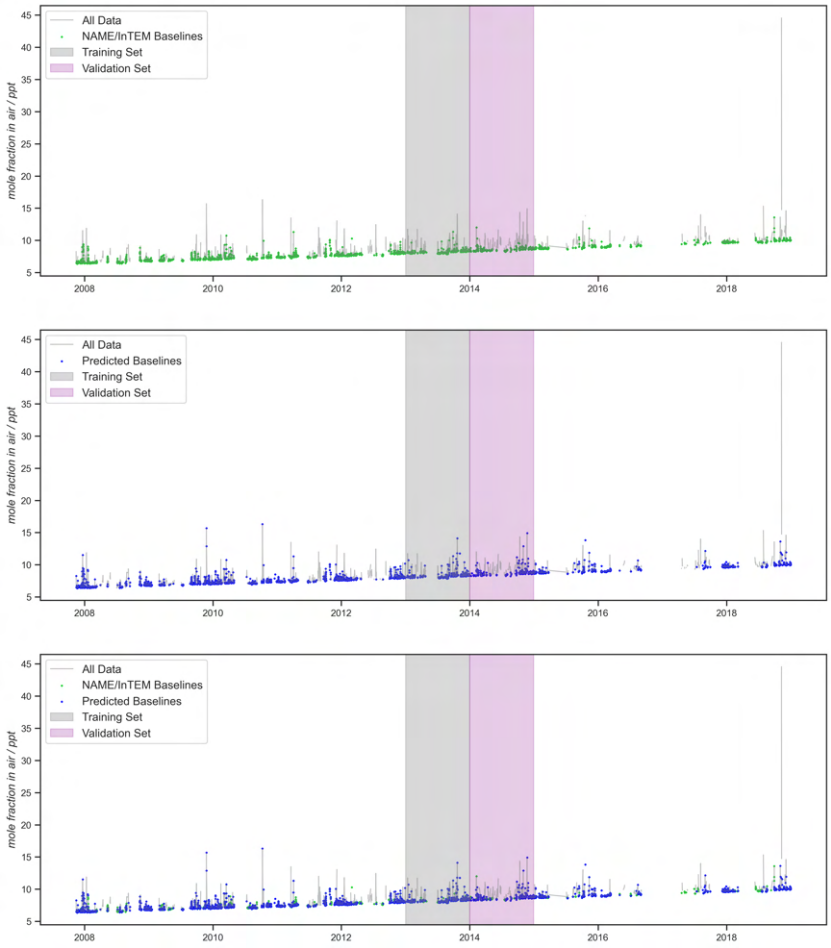


Monthly means

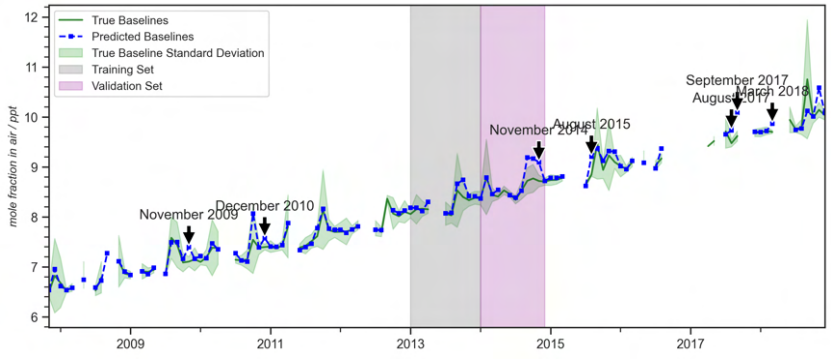


4.6.5 SF_6

Mole fraction time series



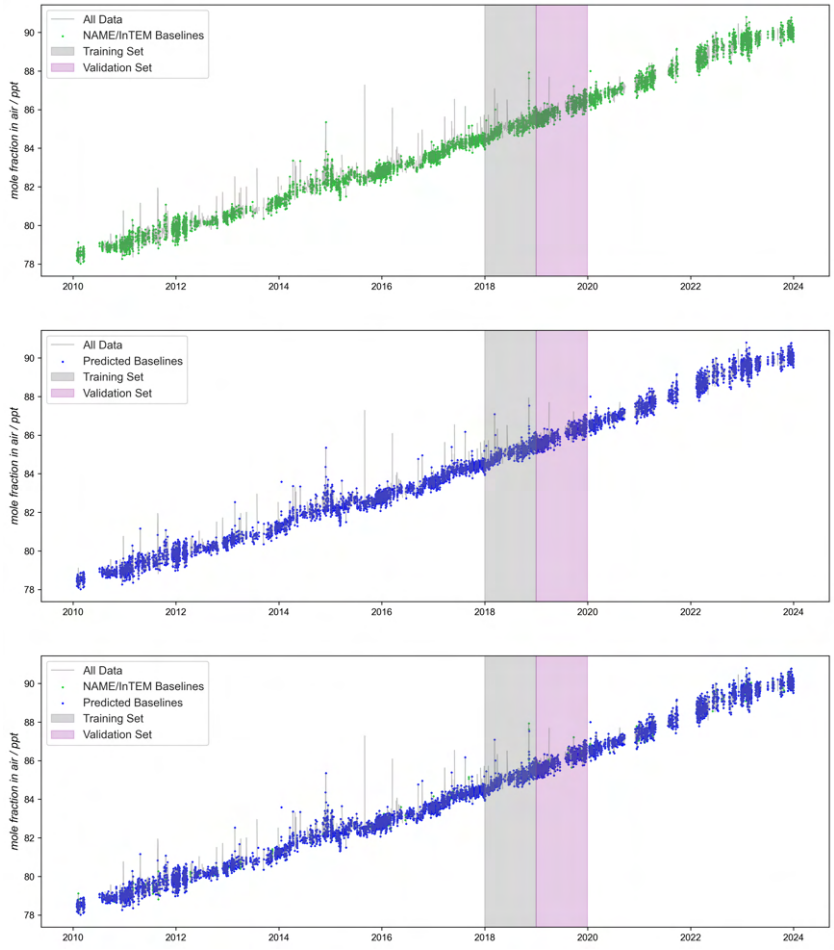
Monthly means



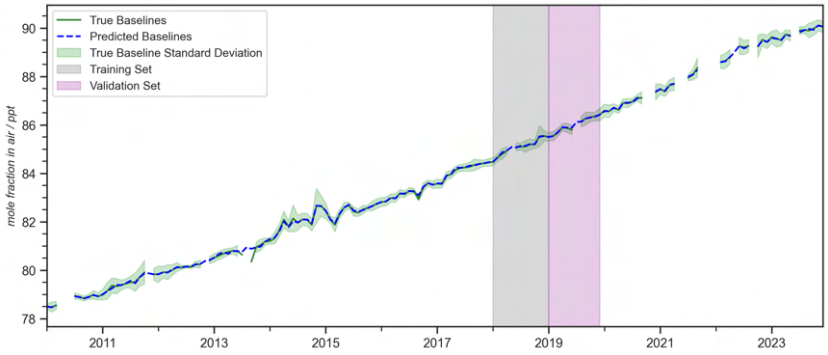
4.7 Jungfraujoch, Switzerland

4.7.1 CF_4

Mole fraction time series

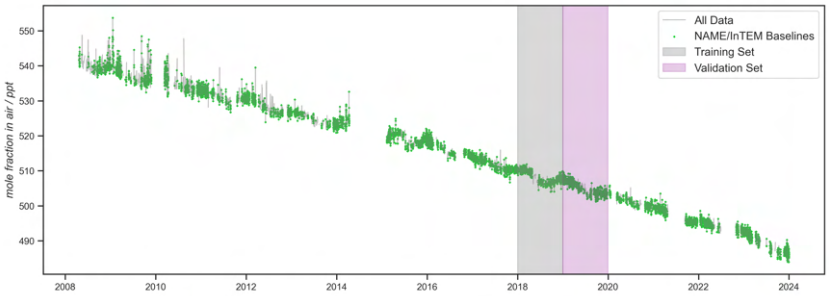


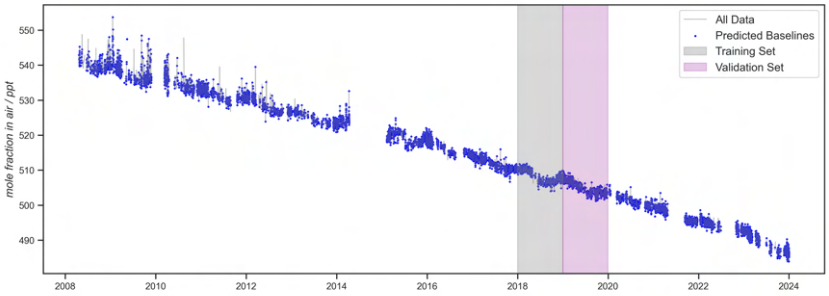
Monthly means

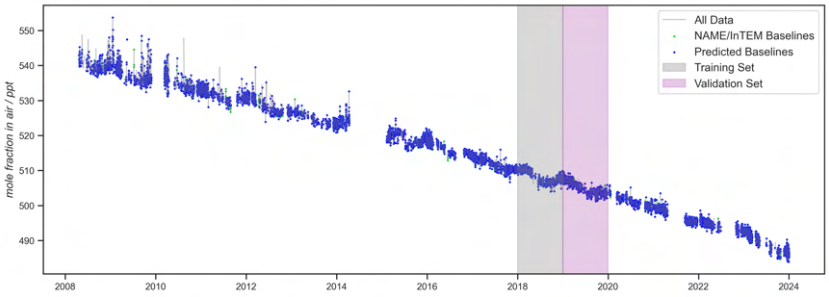


4.7.2 CFC-12

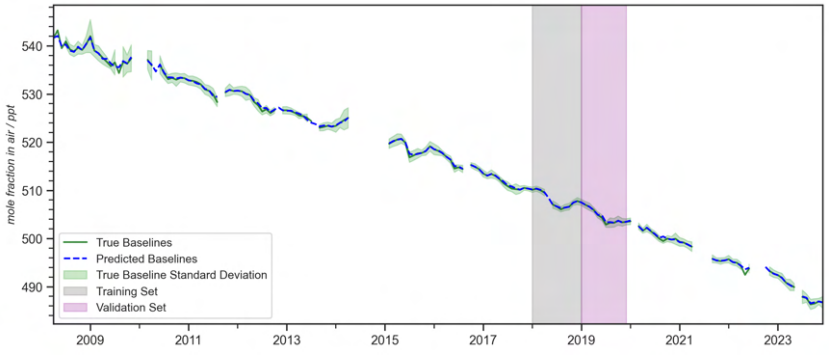
Mole fraction time series





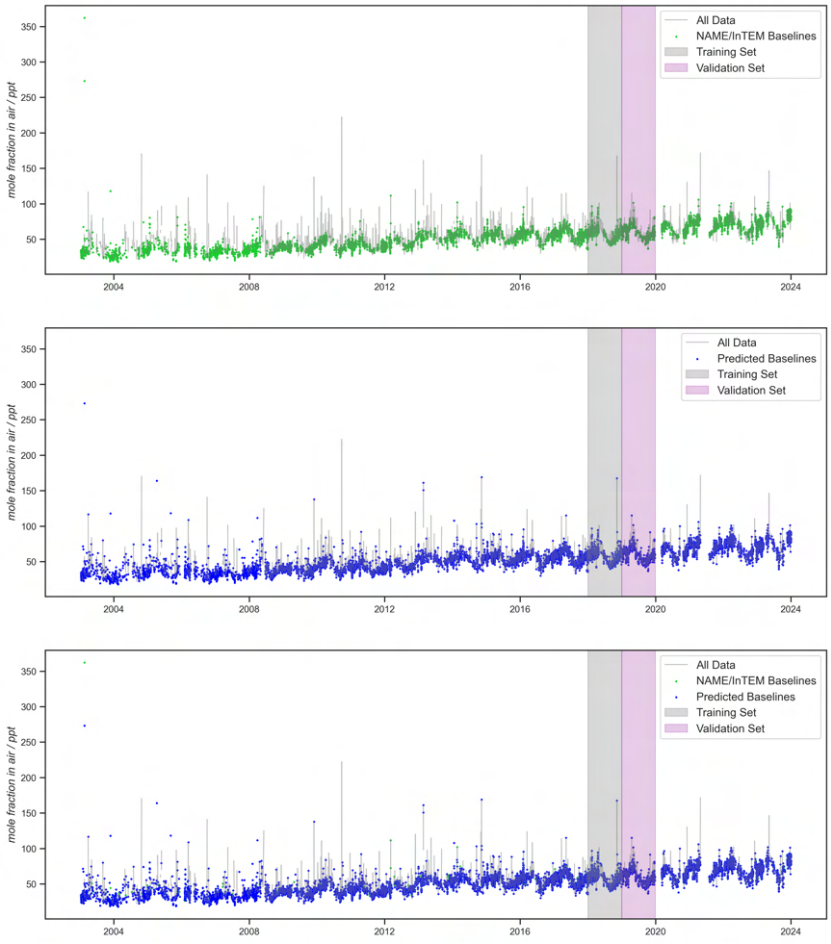


Monthly means

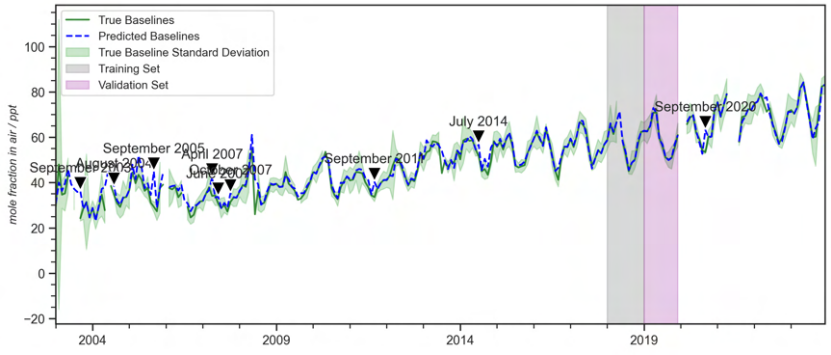


4.7.3 CH_2Cl_2

Mole fraction time series

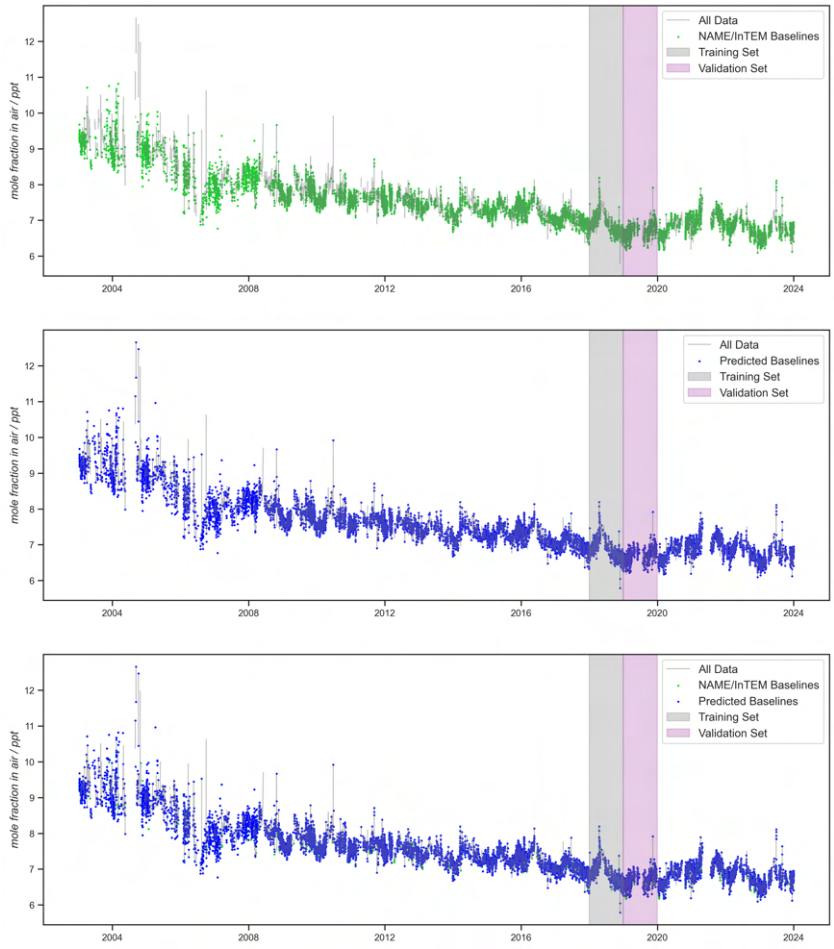


Monthly means

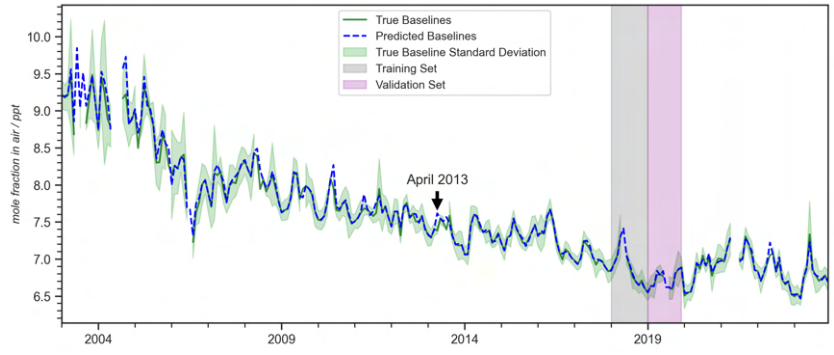


4.7.4 CH_3Br

Mole fraction time series

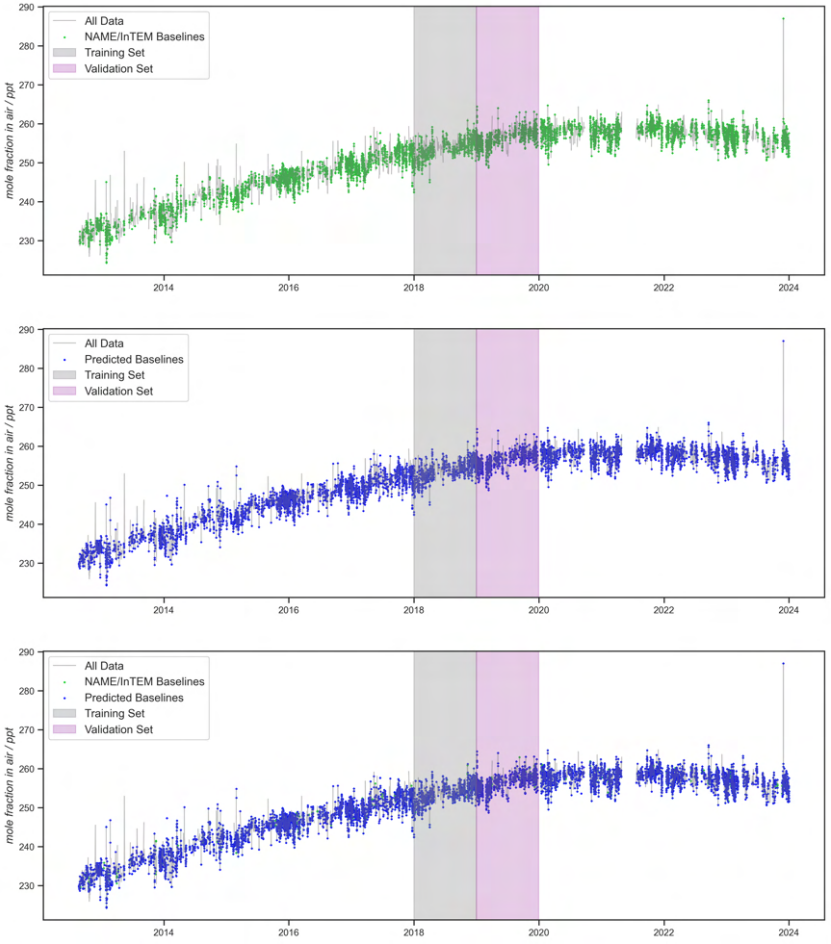


Monthly means

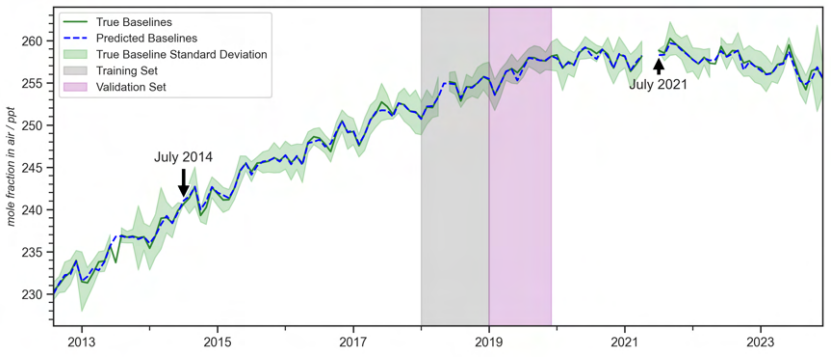


4.7.5 HCFC-22

Mole fraction time series

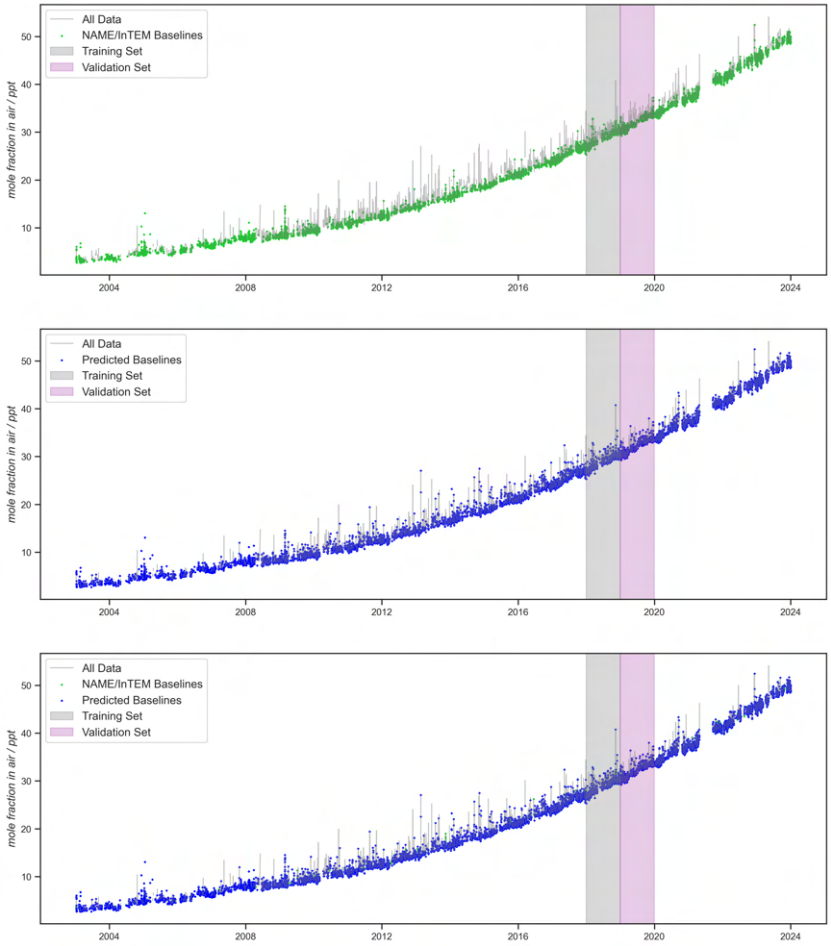


Monthly means

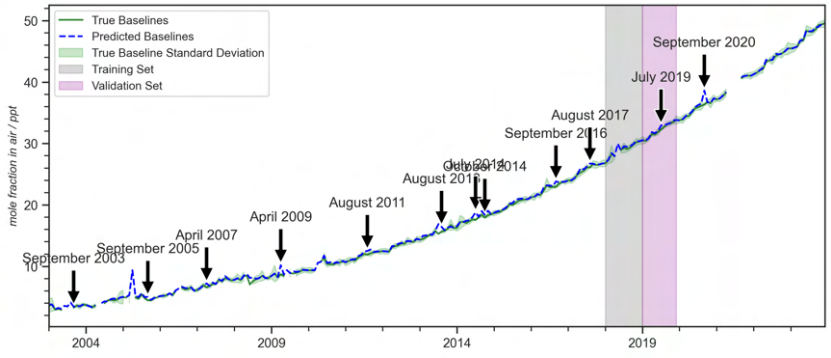


4.7.6 HFC-125

Mole fraction time series

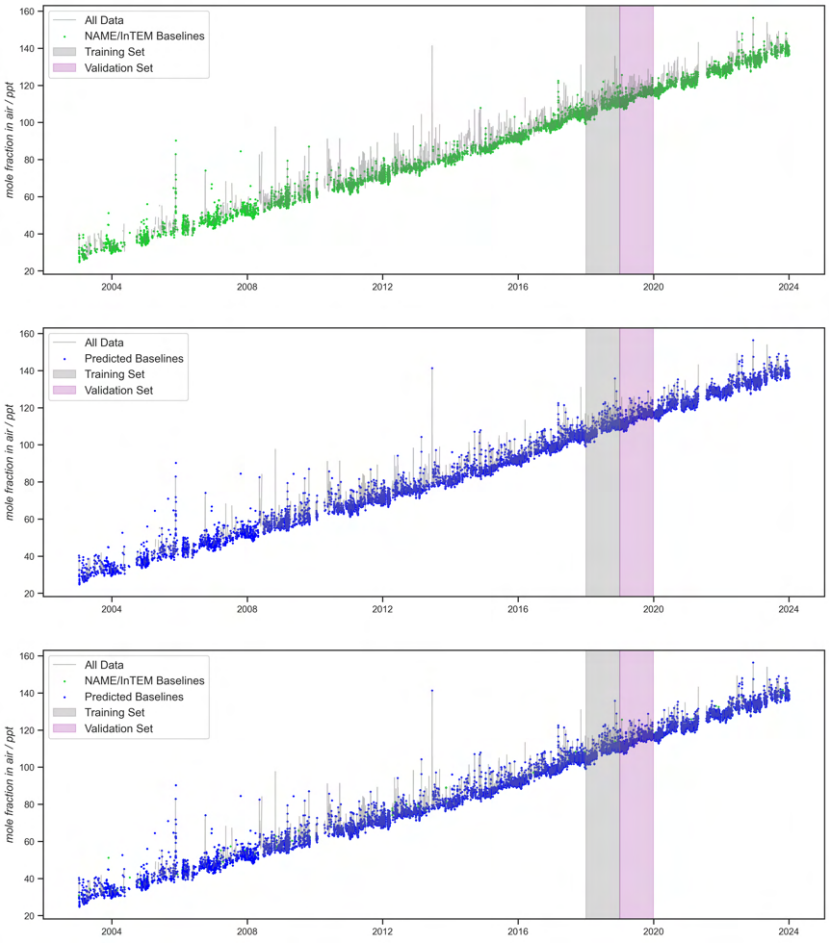


Monthly means

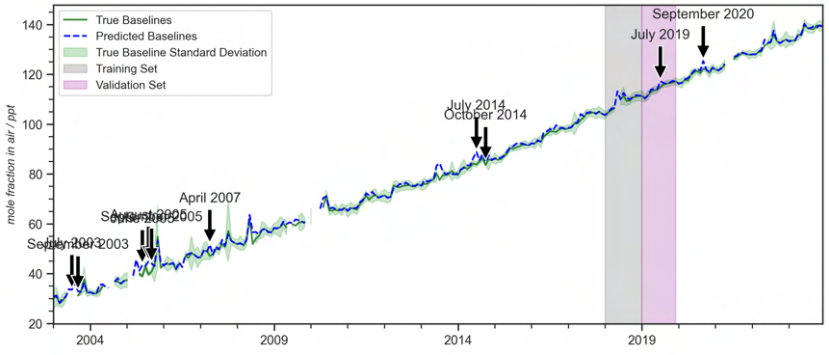


4.7.7 HFC-134a

Mole fraction time series

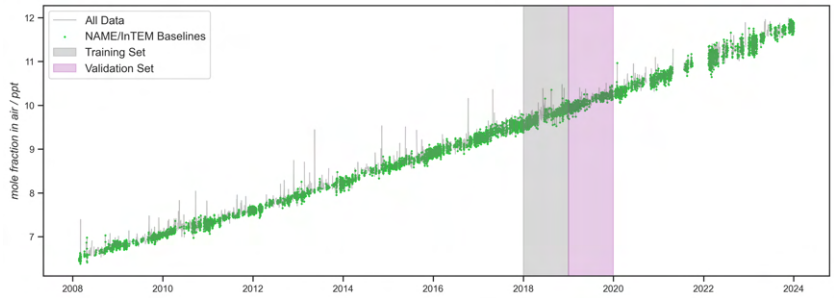


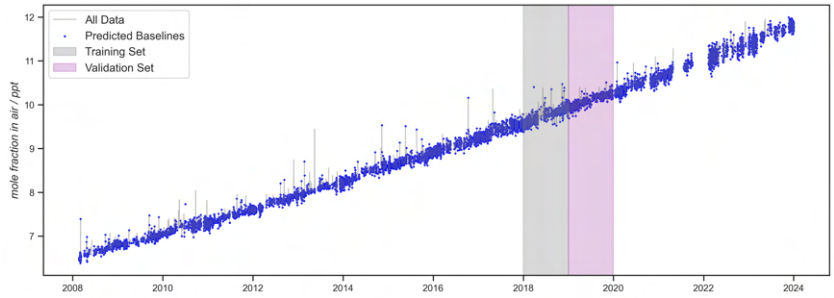
Monthly means

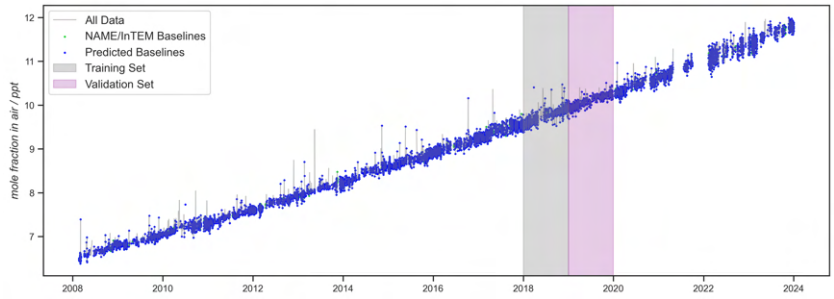


4.7.8 SF_6

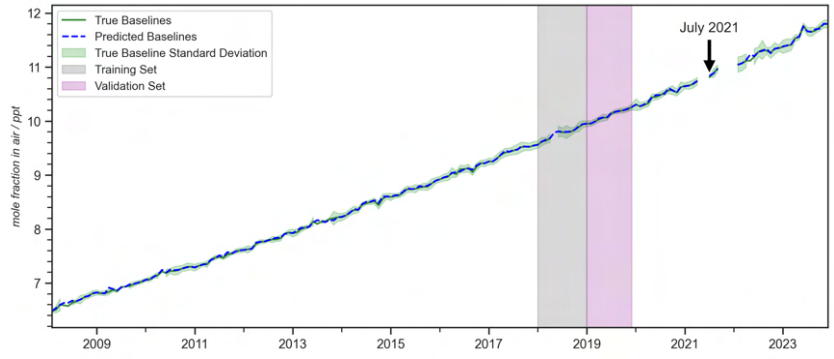
Mole fraction time series







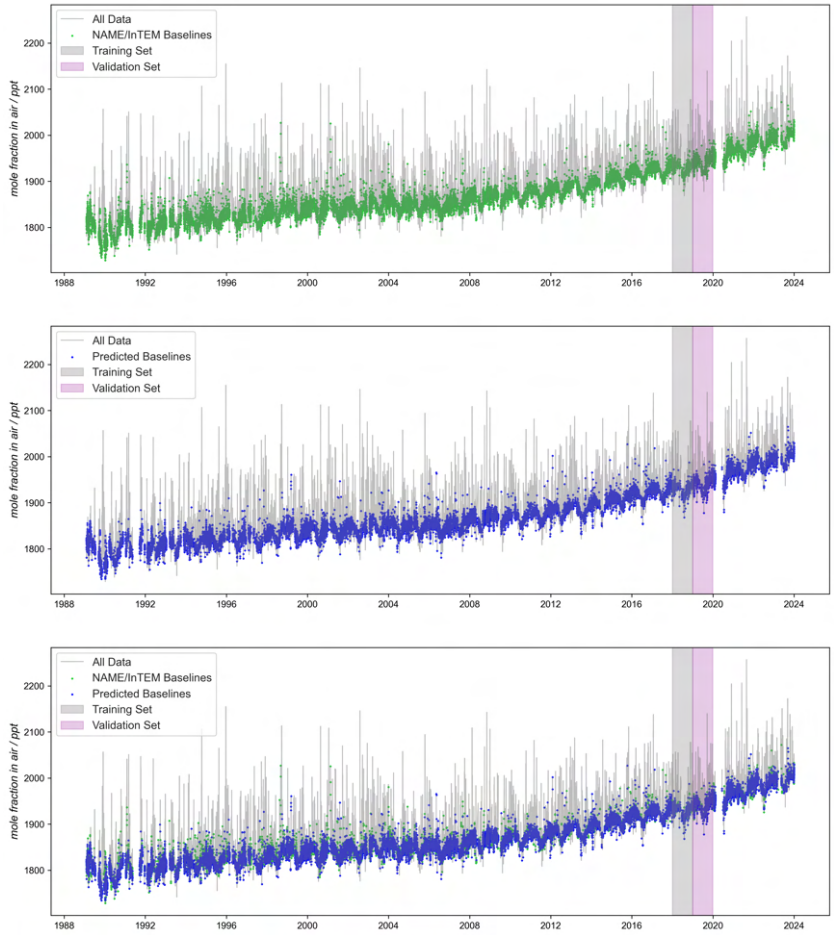
Monthly means



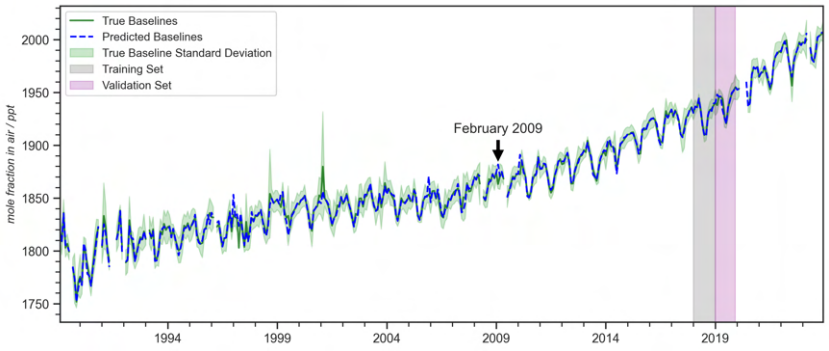
4.8 Mace Head, Ireland

4.8.1 CH_4

Mole fraction time series

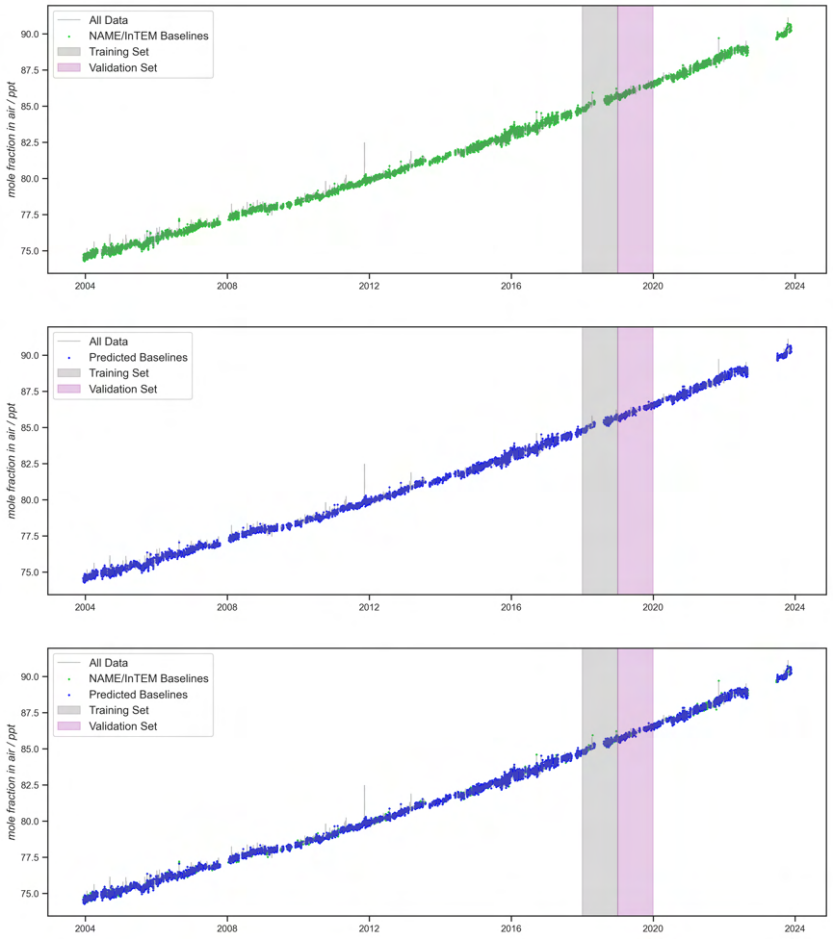


Monthly means

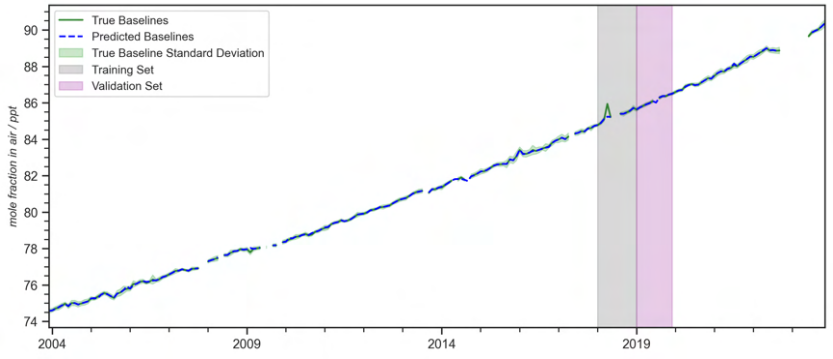


4.8.2 CF_4

Mole fraction time series

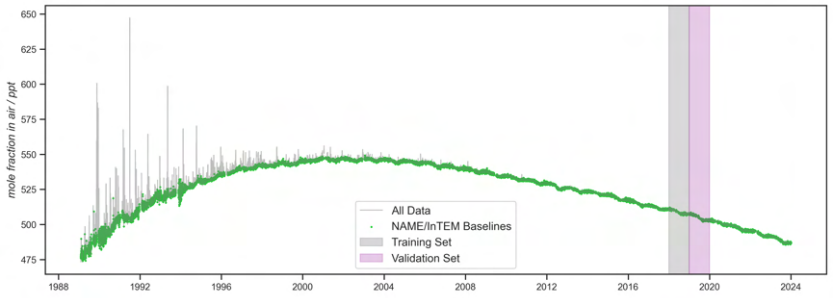


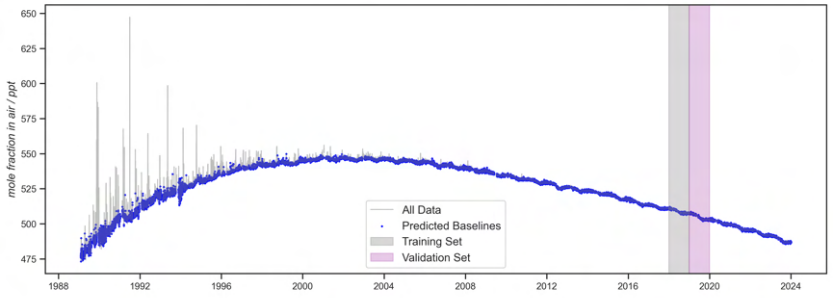
Monthly means

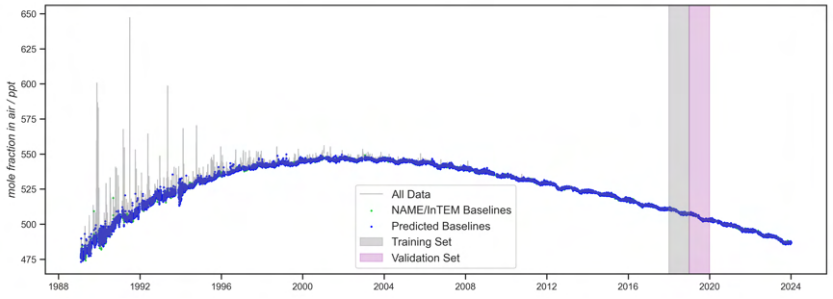


4.8.3 CFC-12

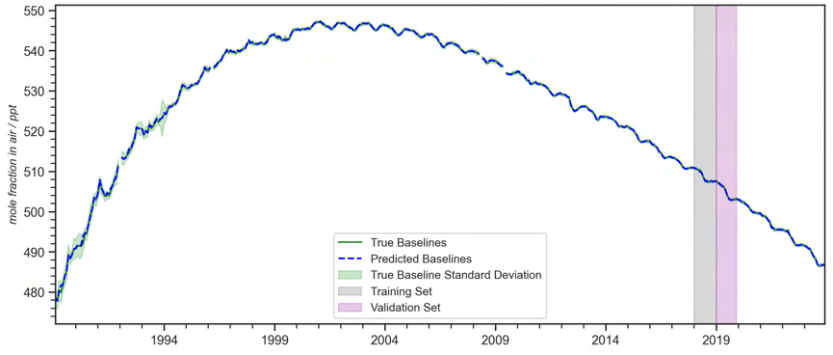
Mole fraction time series





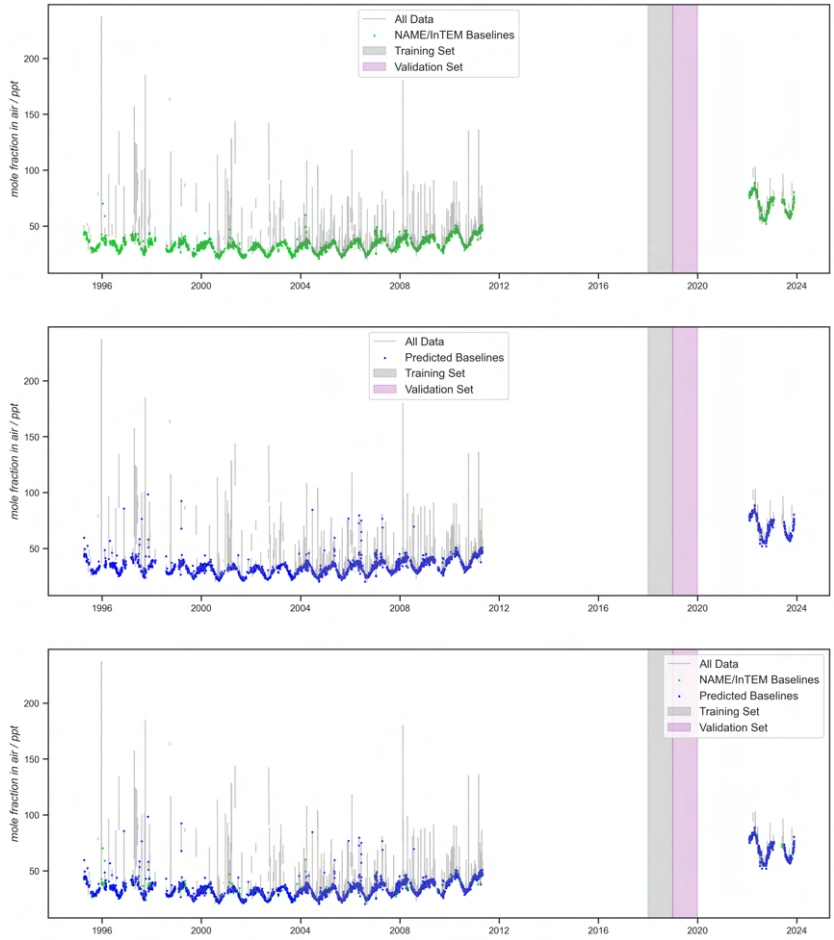


Monthly means

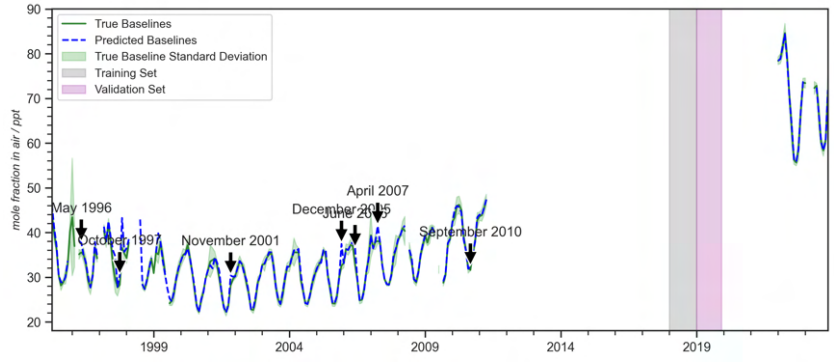


4.8.4 CH_2Cl_2

Mole fraction time series

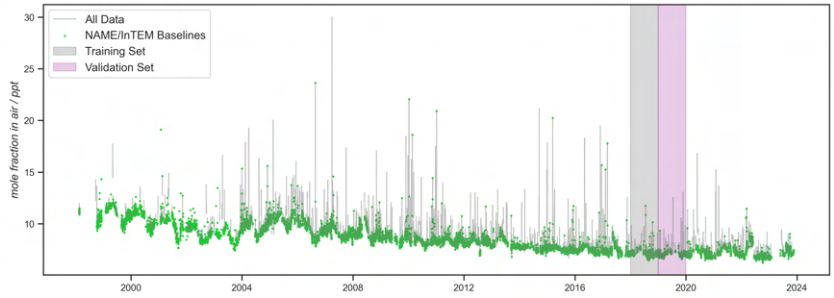


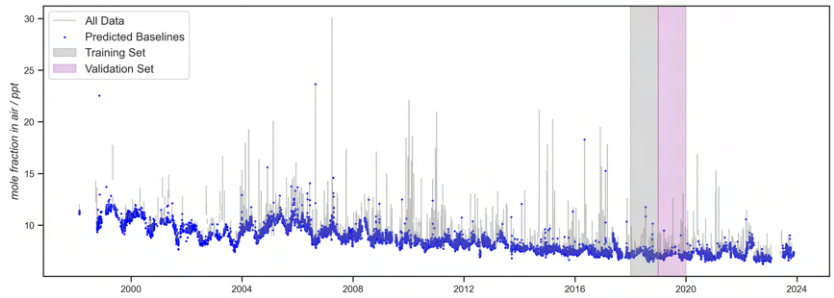
Monthly means

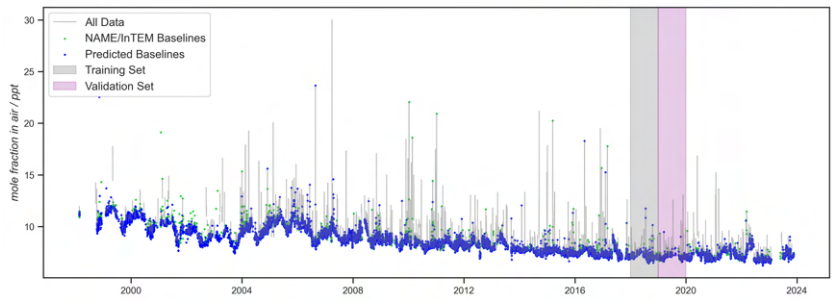


4.8.5 CH_3Br

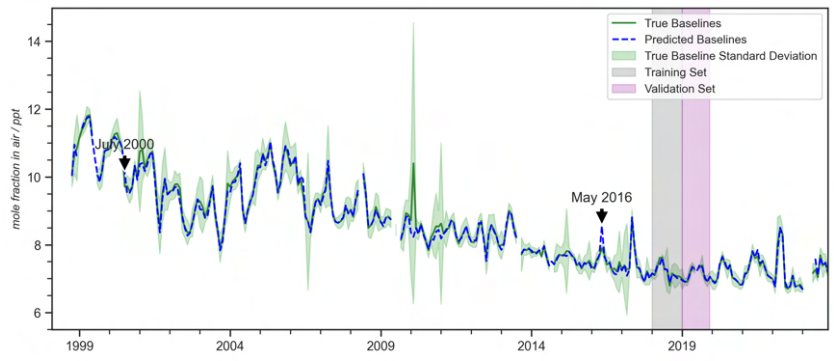
Mole fraction time series





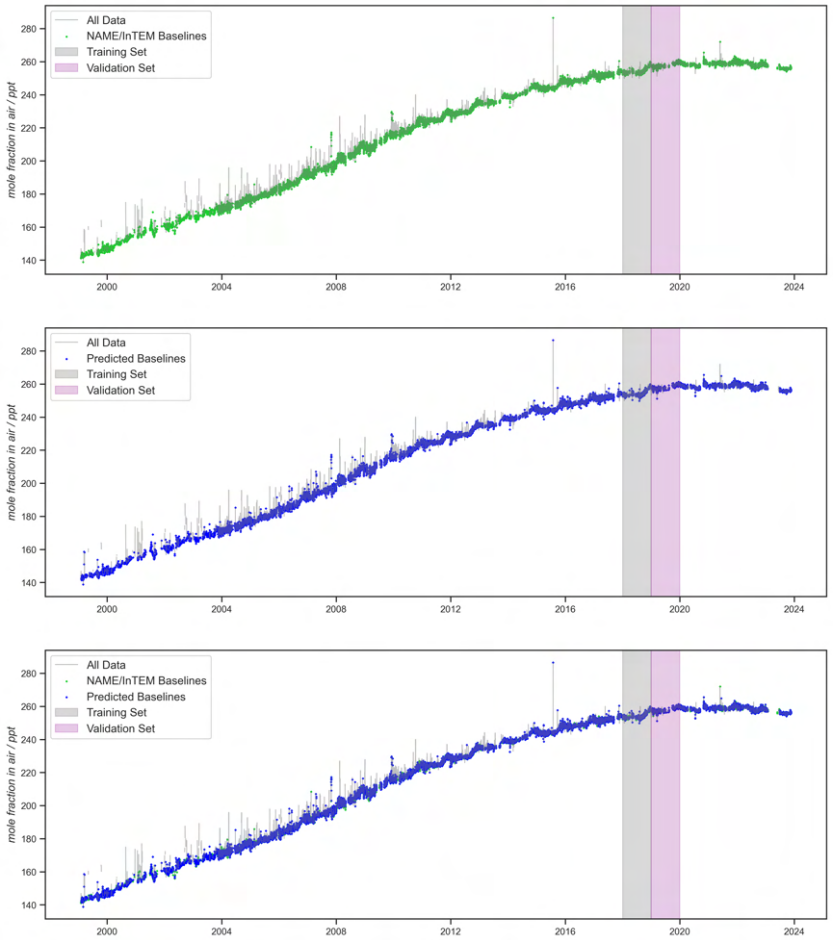


Monthly means

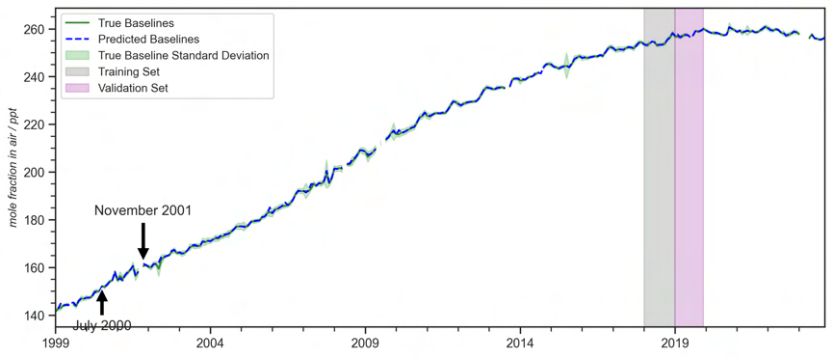


4.8.6 HCFC-22

Mole fraction time series

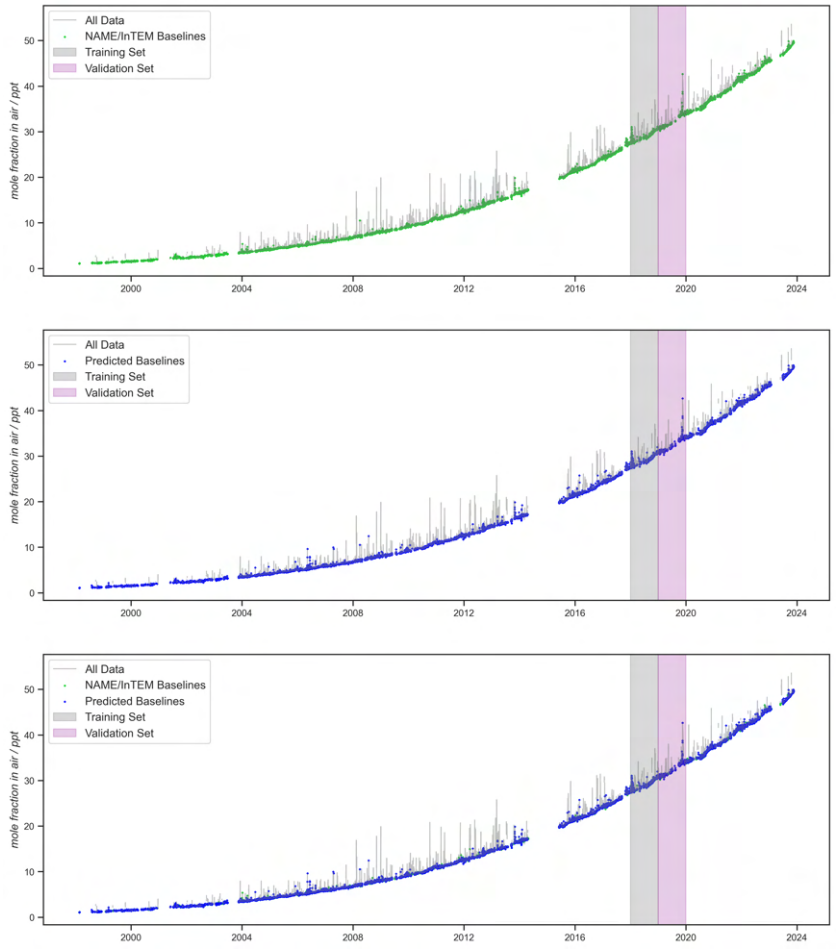


Monthly means

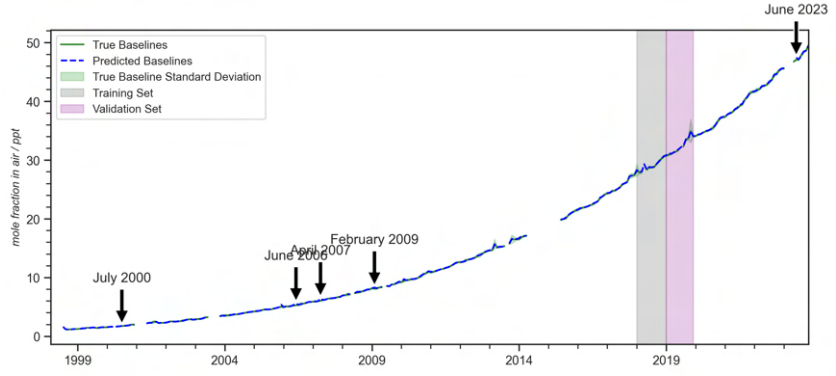


4.8.7 HFC-125

Mole fraction time series

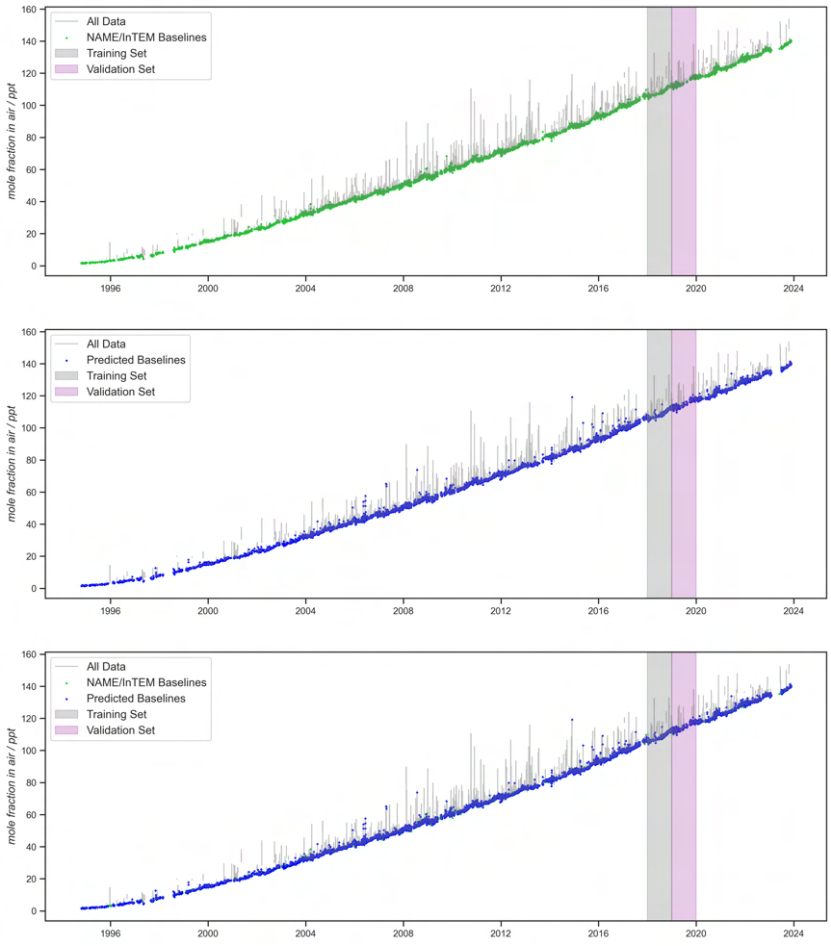


Monthly means

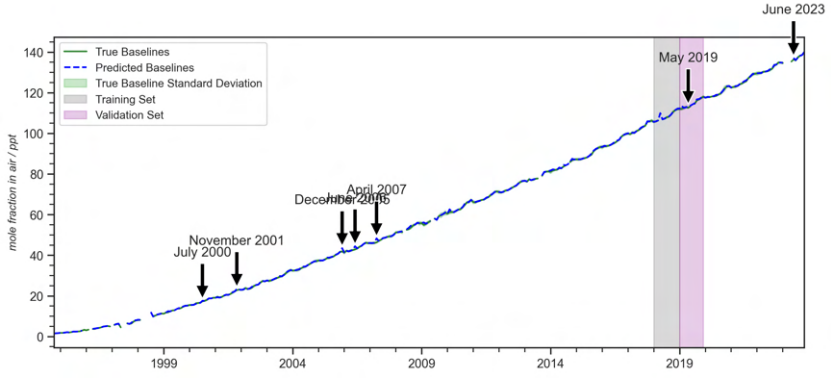


4.8.8 HFC-134a

Mole fraction time series

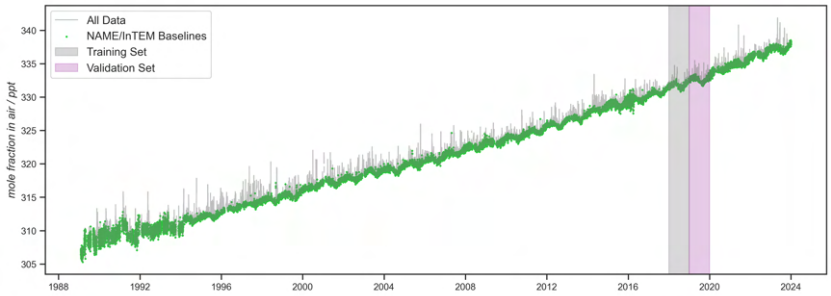


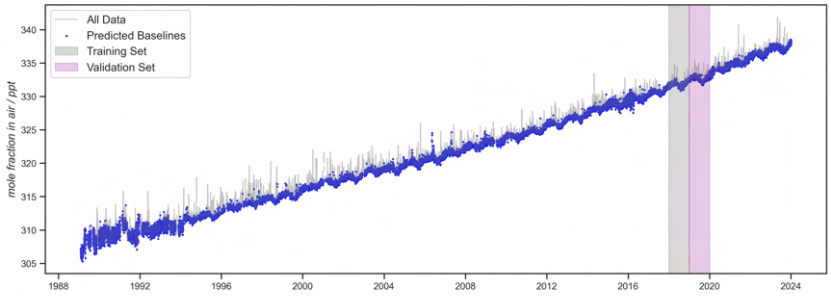
Monthly means

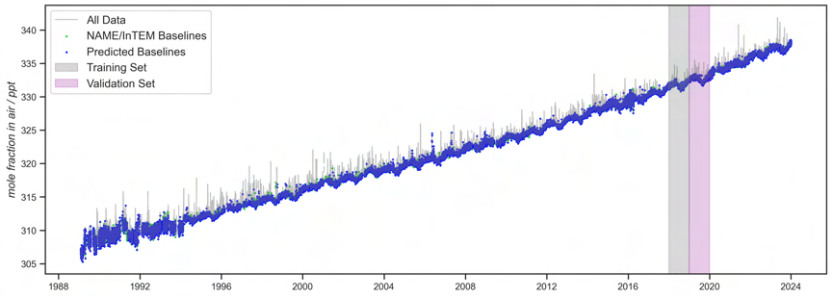


4.8.9 N_2O

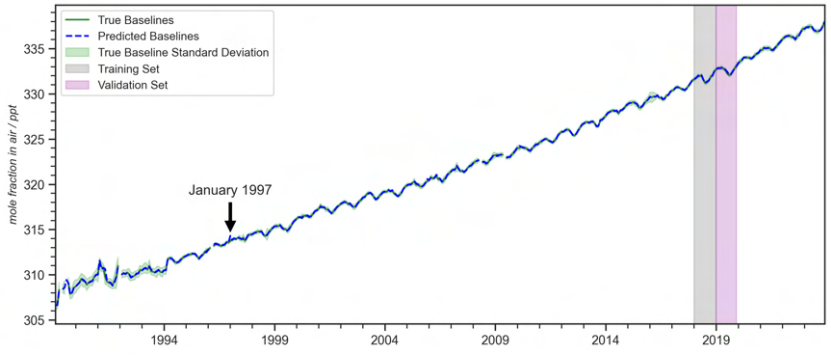
Mole fraction time series





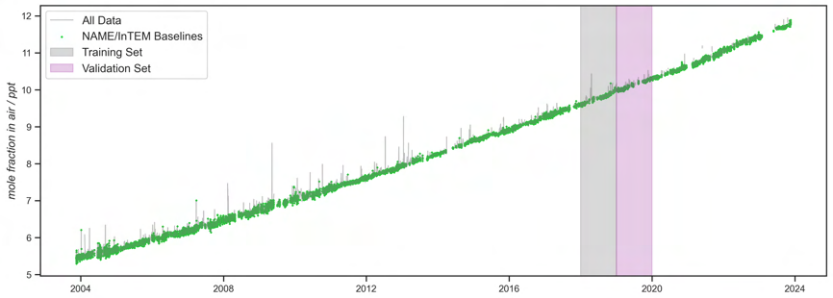


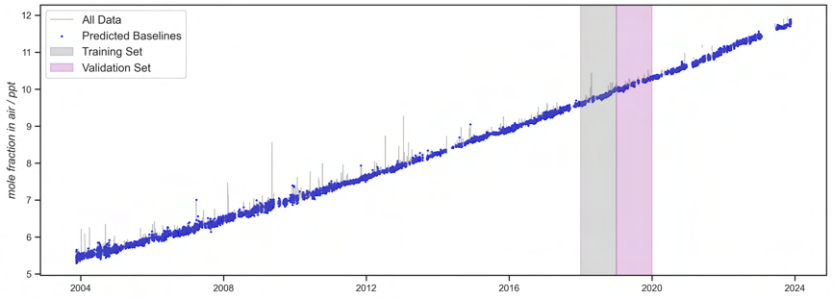
Monthly means

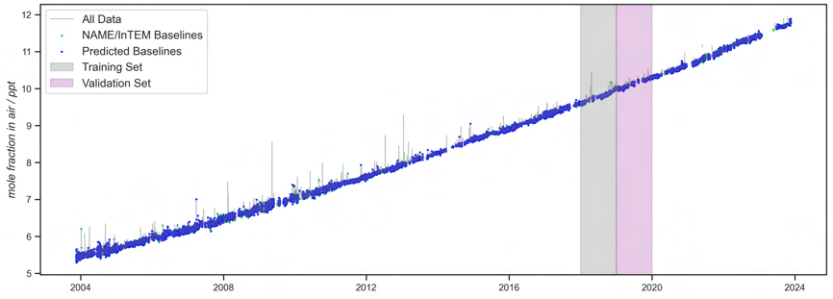


4.8.10 SF_6

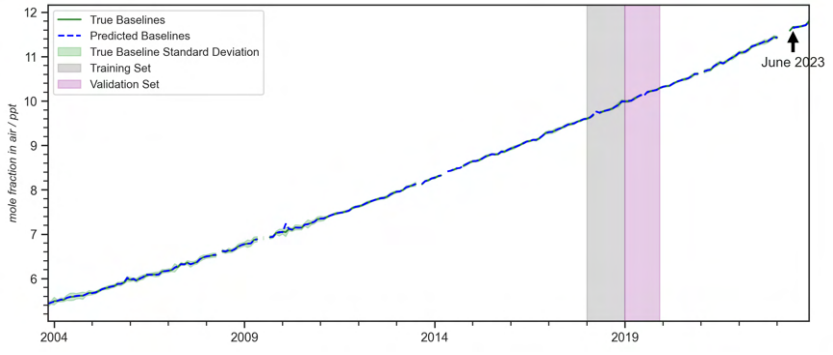
Mole fraction time series







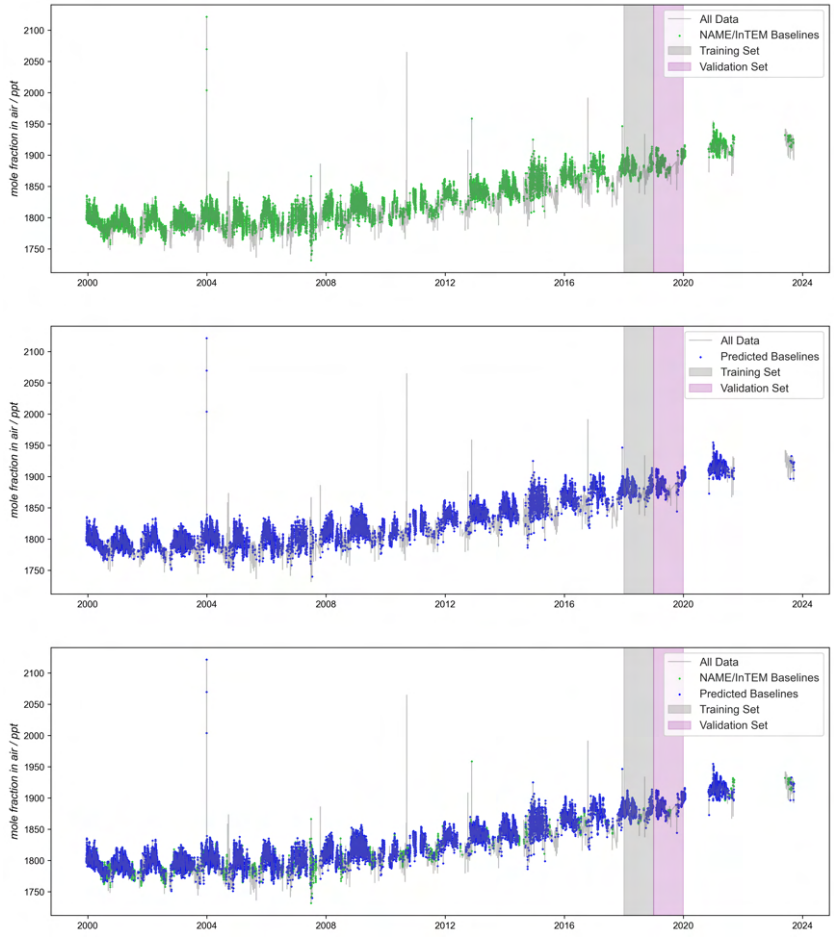
Monthly means



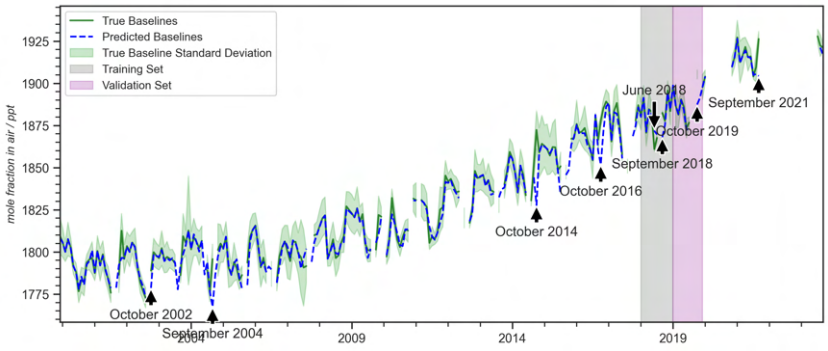
4.9 Ragged Point, Barbados

4.9.1 CH_4

Mole fraction time series

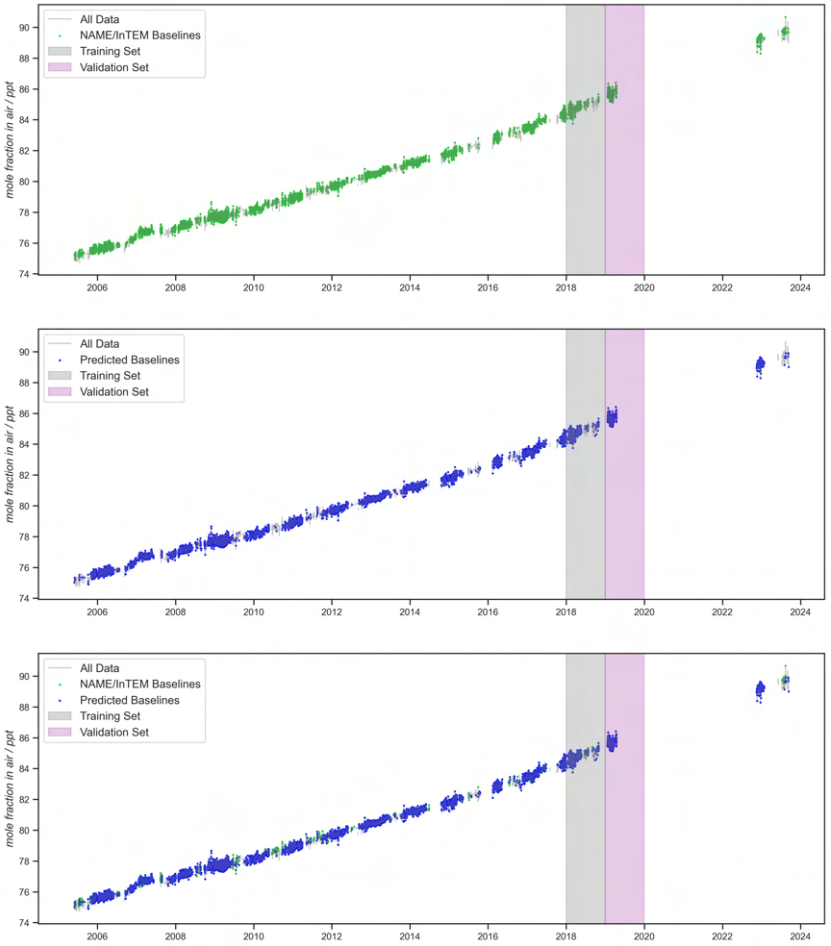


Monthly means

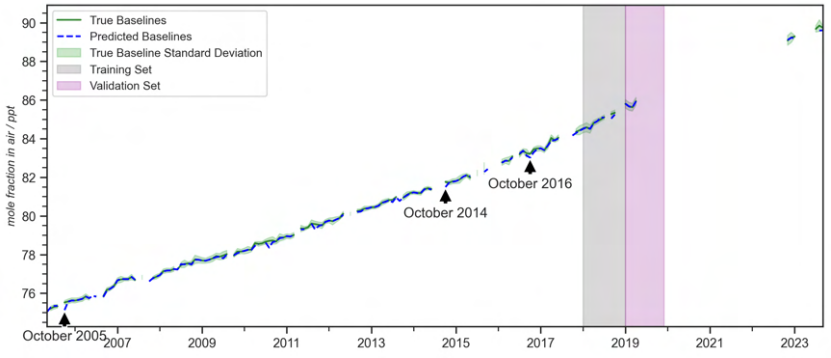


4.9.2 CF_4

Mole fraction time series

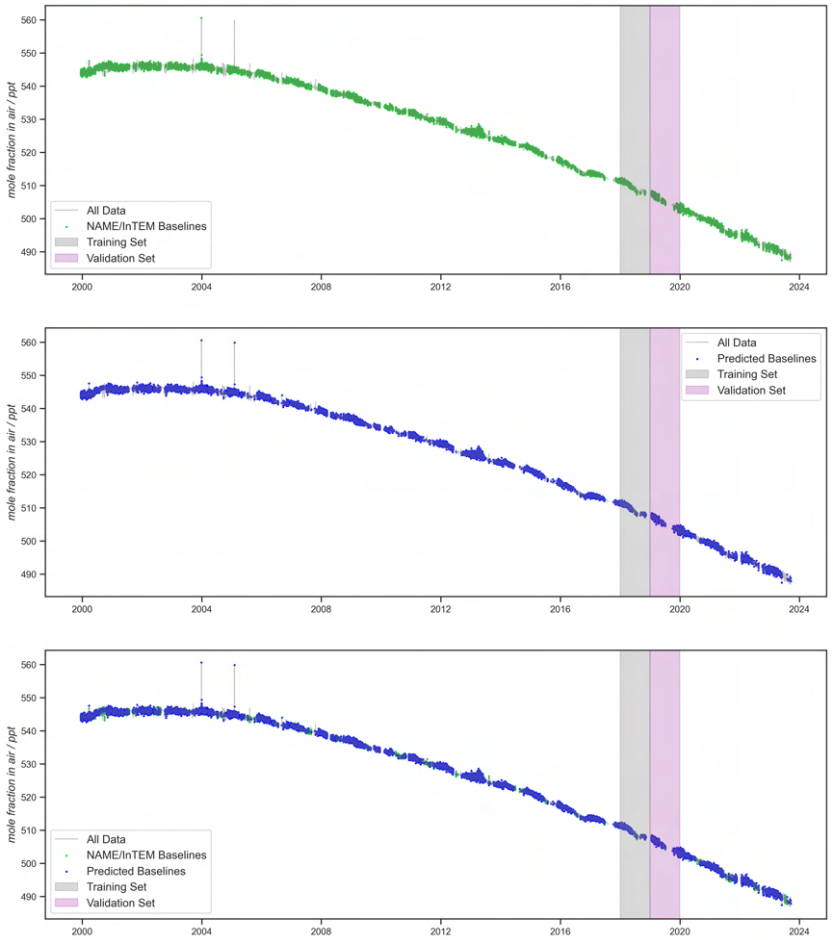


Monthly means

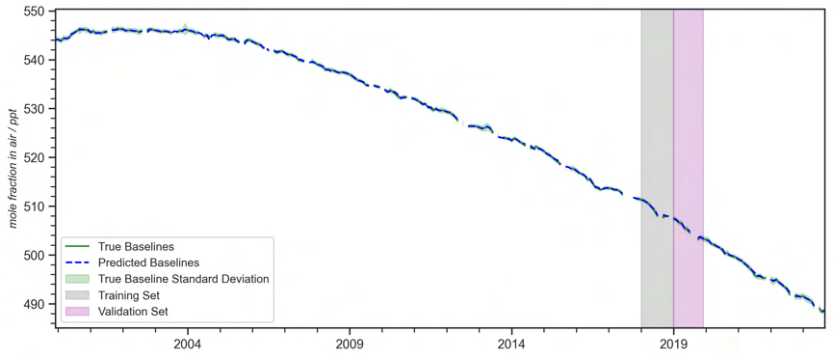


4.9.3 CFC-12

Mole fraction time series

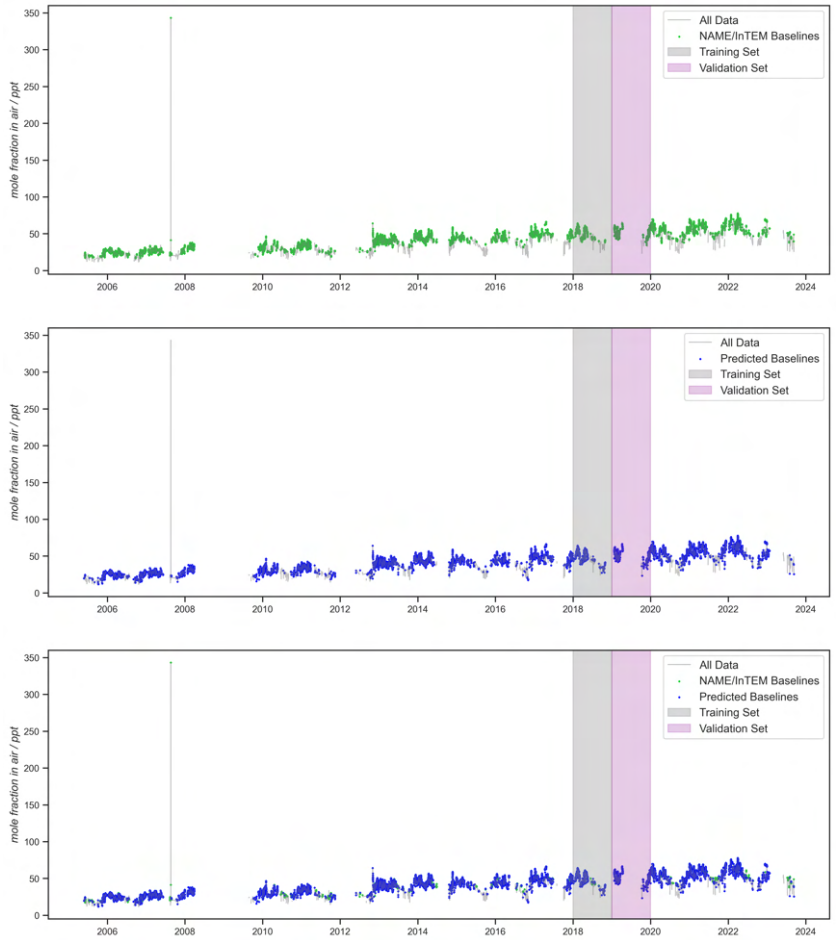


Monthly means

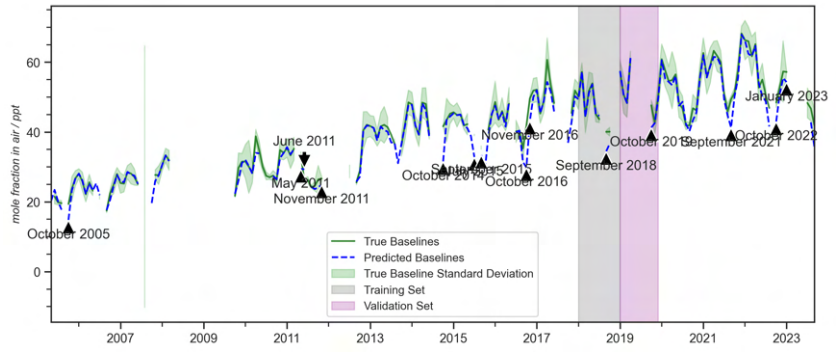


4.9.4 CH_2Cl_2

Mole fraction time series

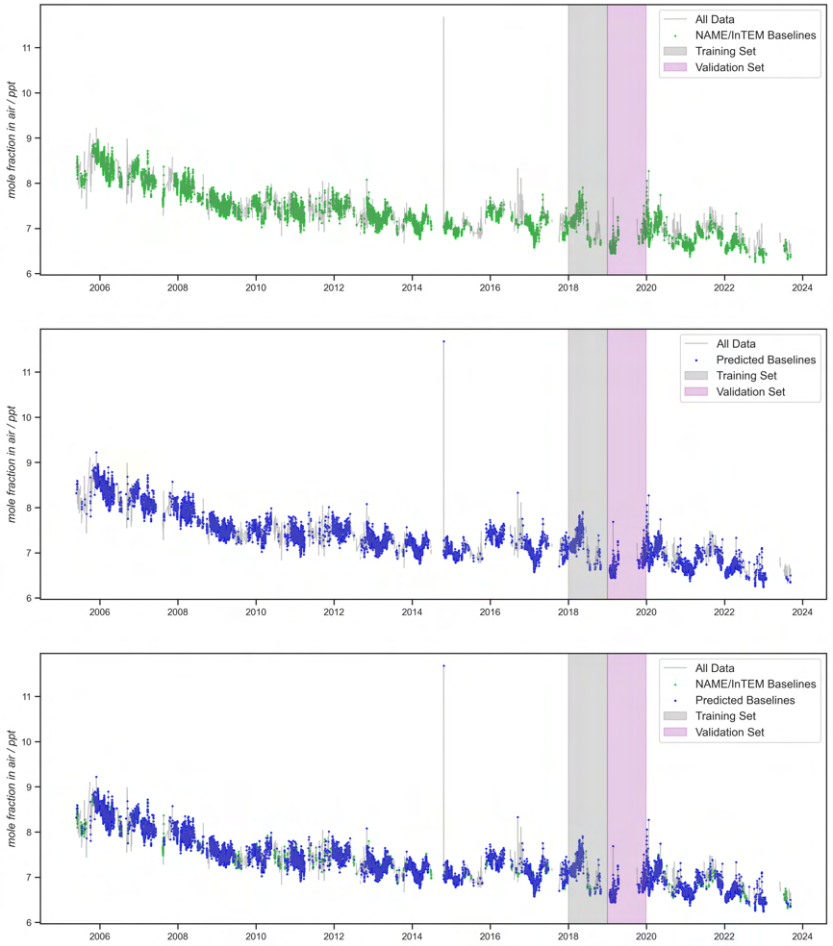


Monthly means

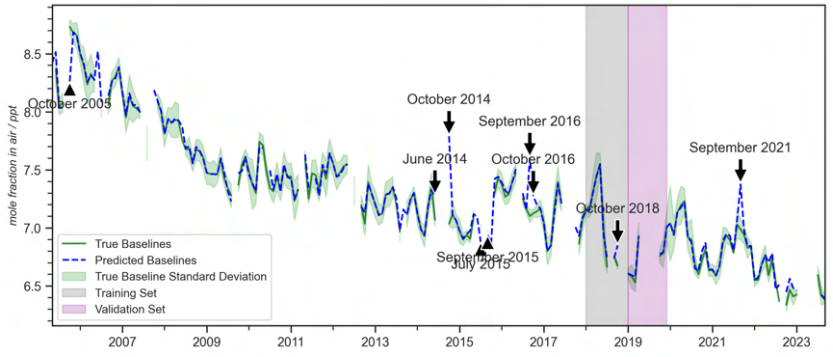


4.9.5 CH_3Br

Mole fraction time series

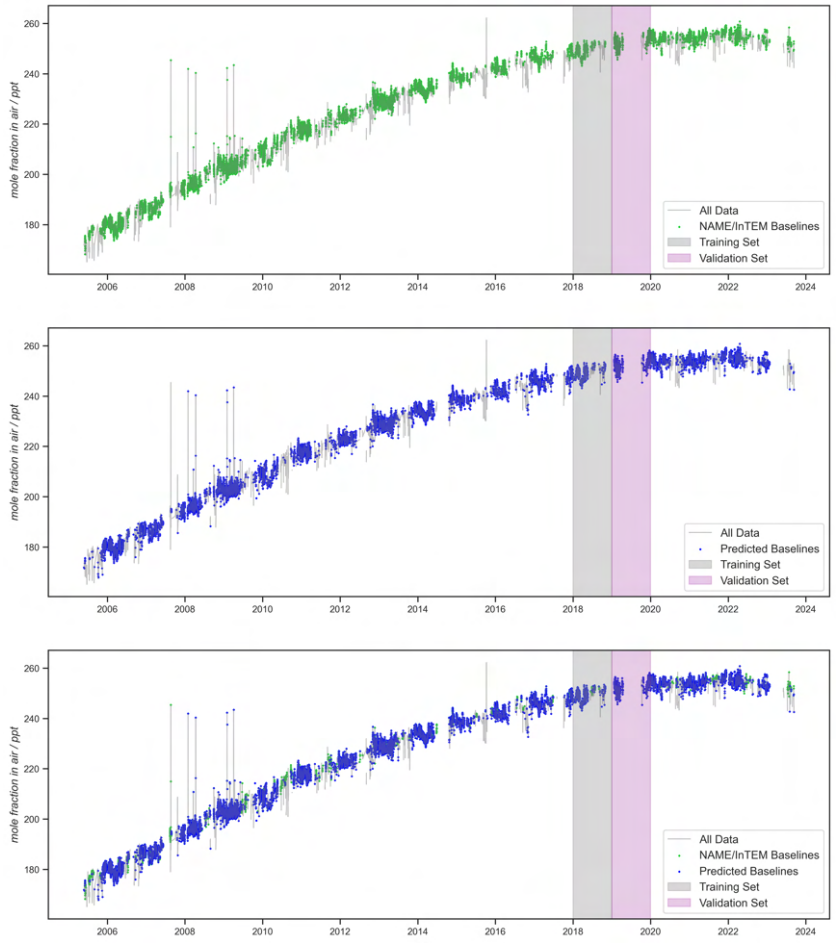


Monthly means

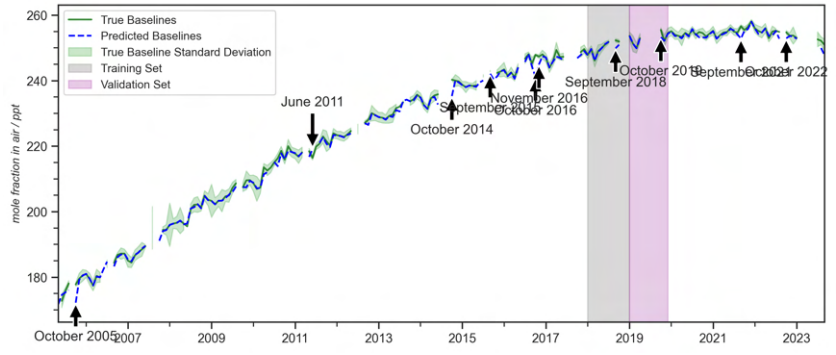


4.9.6 HCFC-22

Mole fraction time series

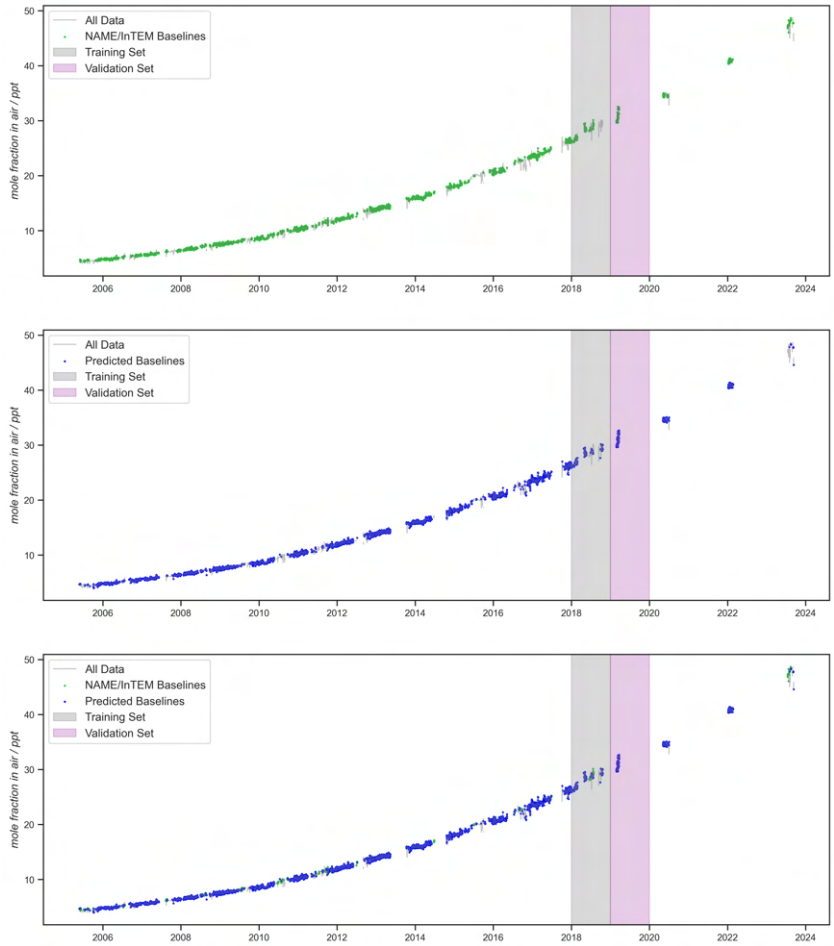


Monthly means

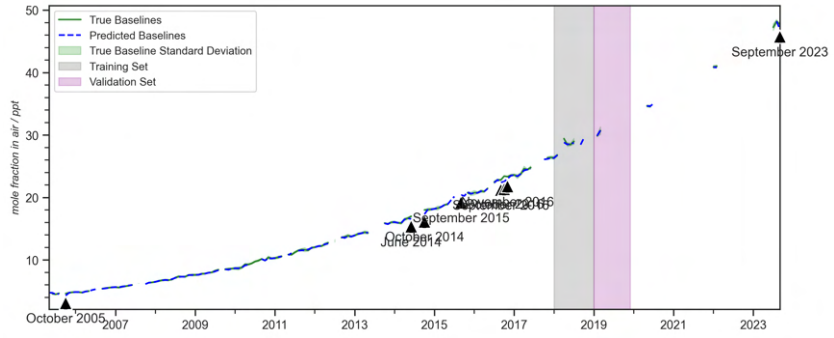


4.9.7 HFC-125

Mole fraction time series

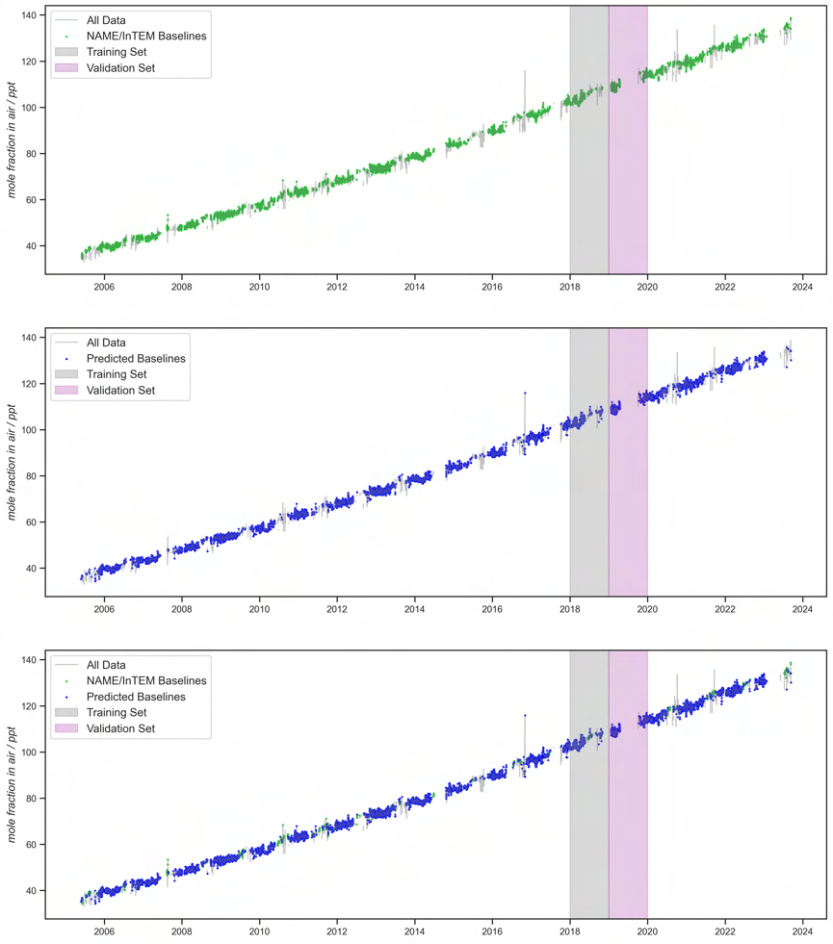


Monthly means

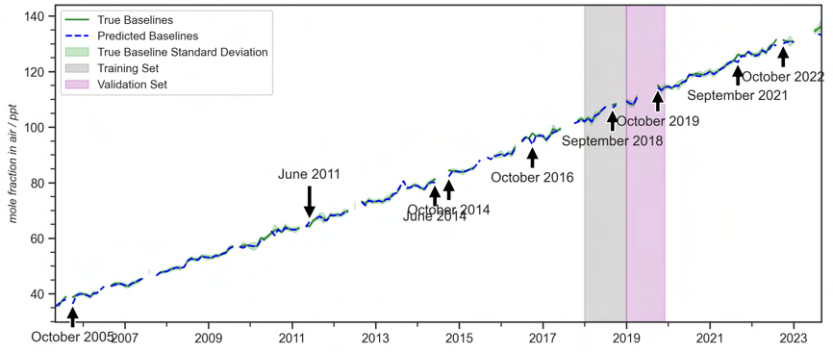


4.9.8 HFC-134a

Mole fraction time series

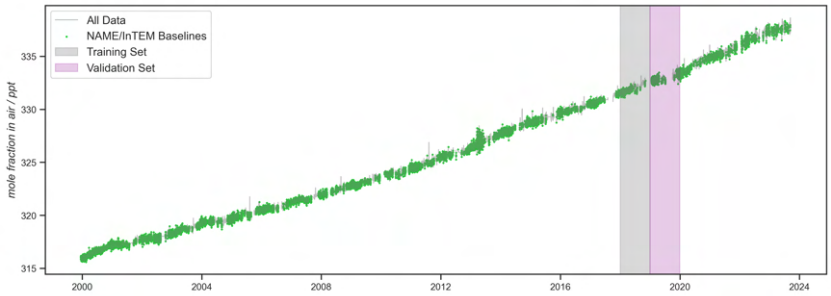


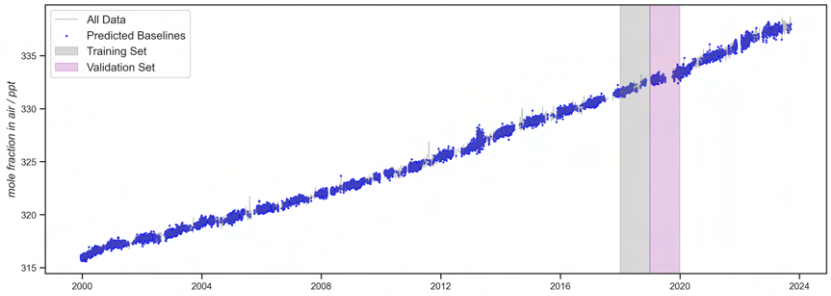
Monthly means

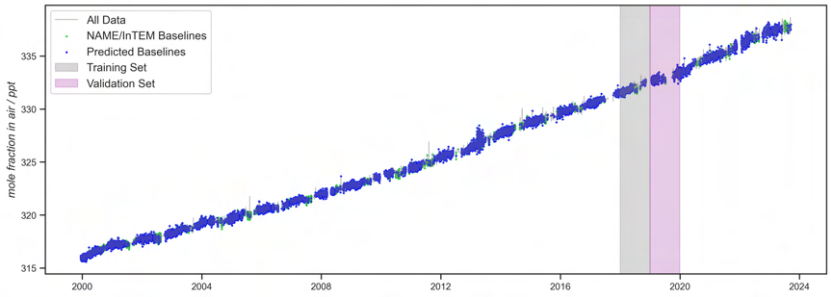


4.9.9 N_2O

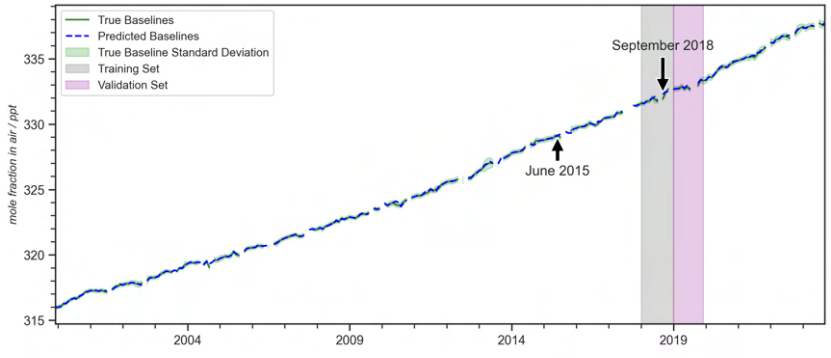
Mole fraction time series





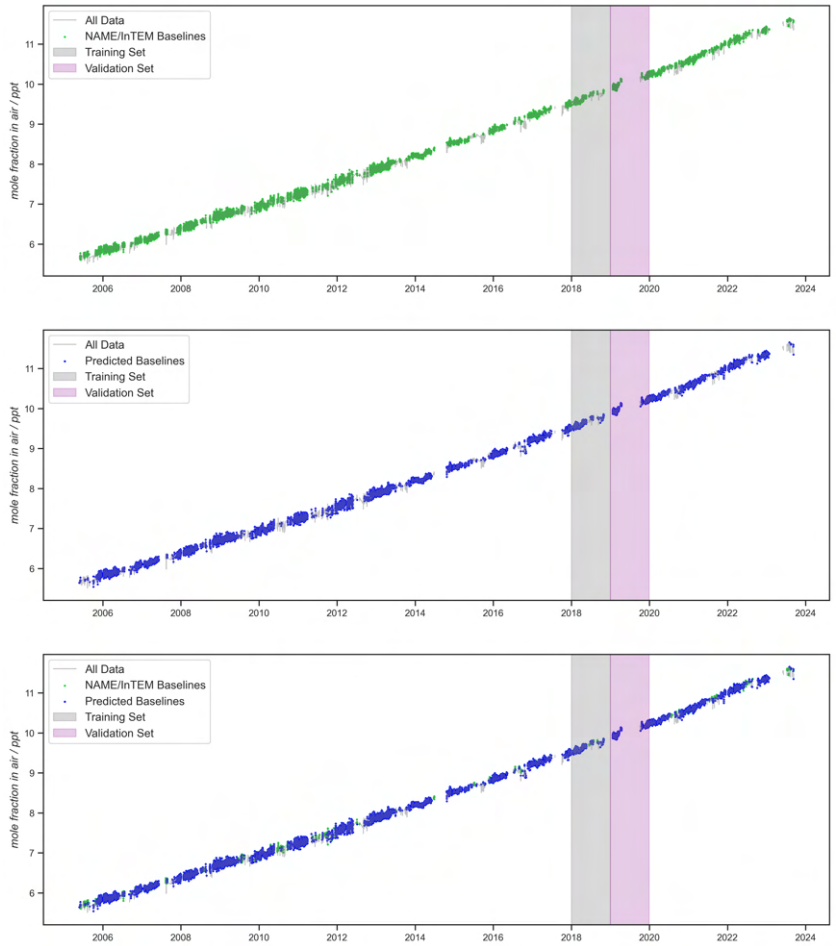


Monthly means

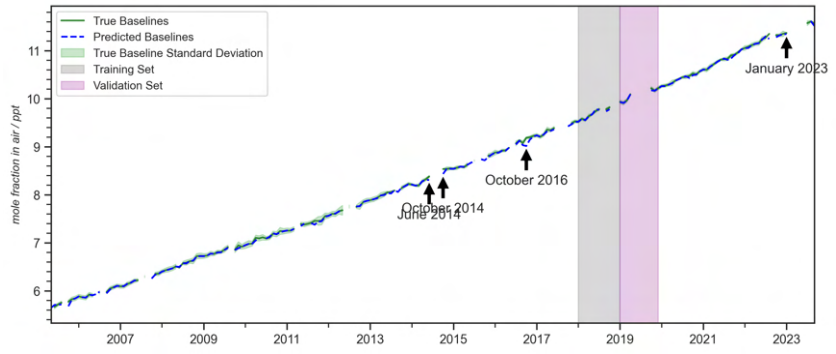


4.9.10 SF_6

Mole fraction time series



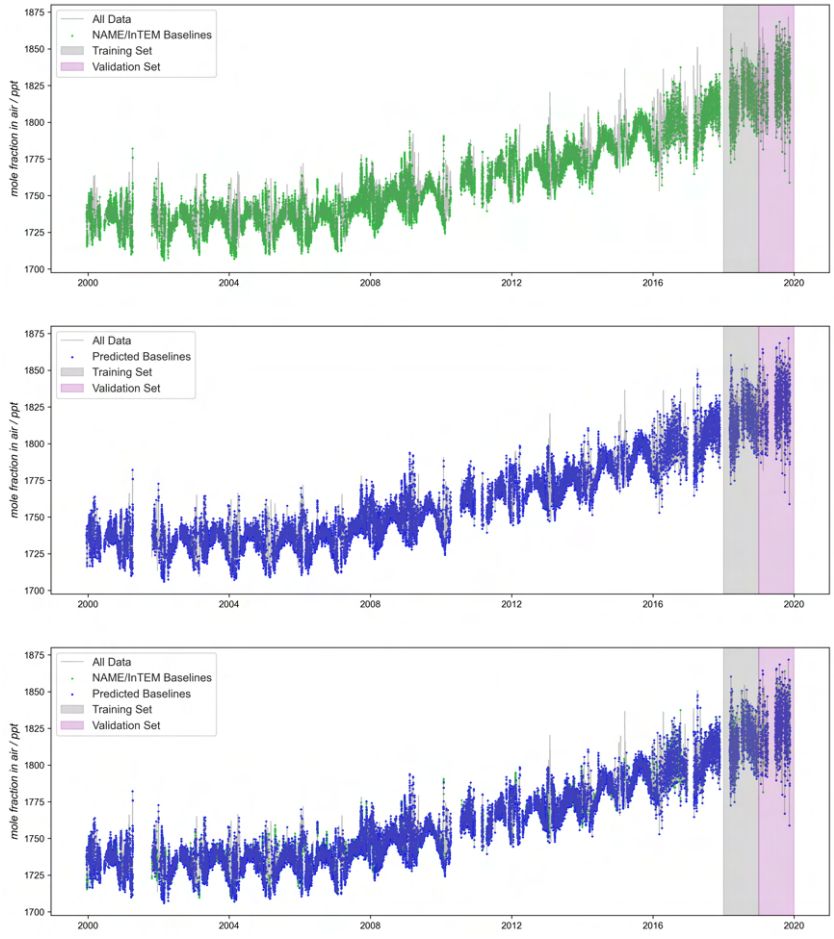
Monthly means



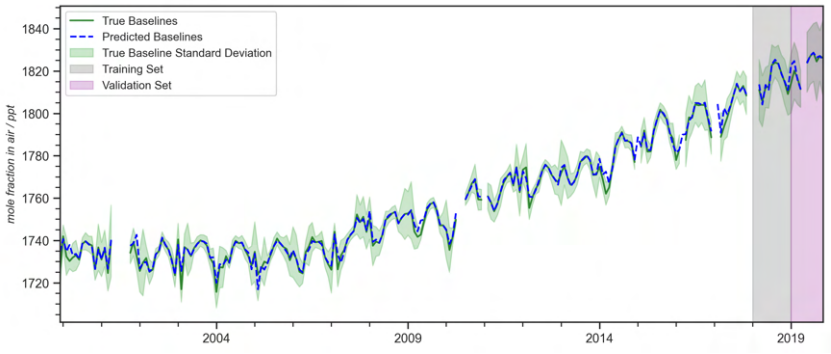
4.10 Cape Matatula, American Samoa

4.10.1 CH_4

Mole fraction time series

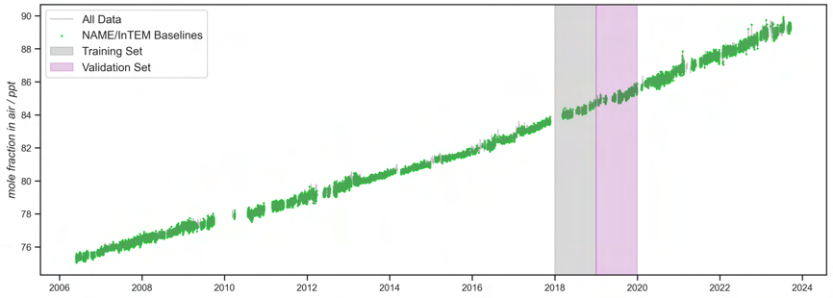


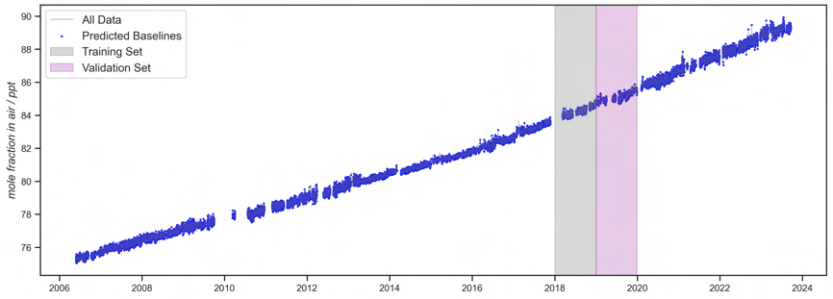
Monthly means

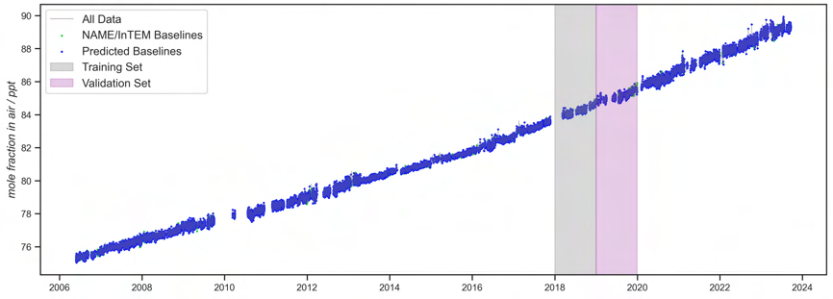


4.10.2 CF_4

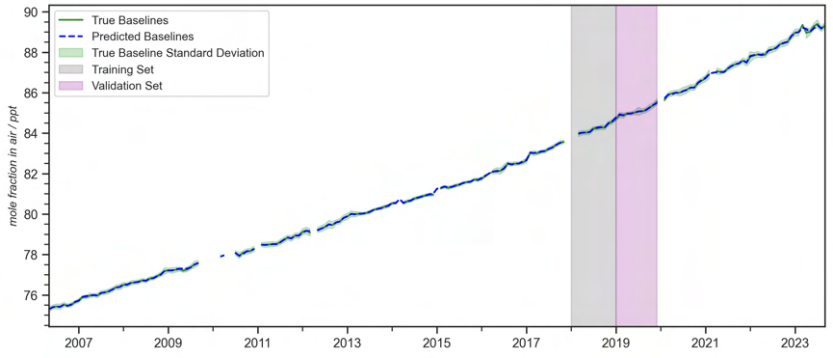
Mole fraction time series





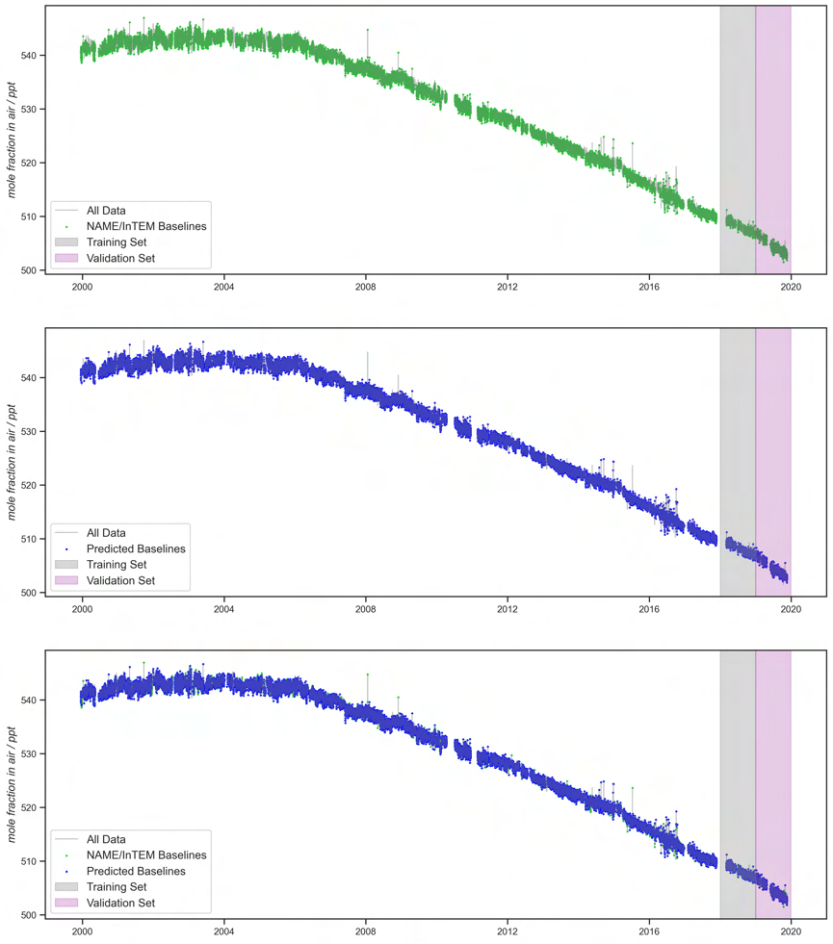


Monthly means

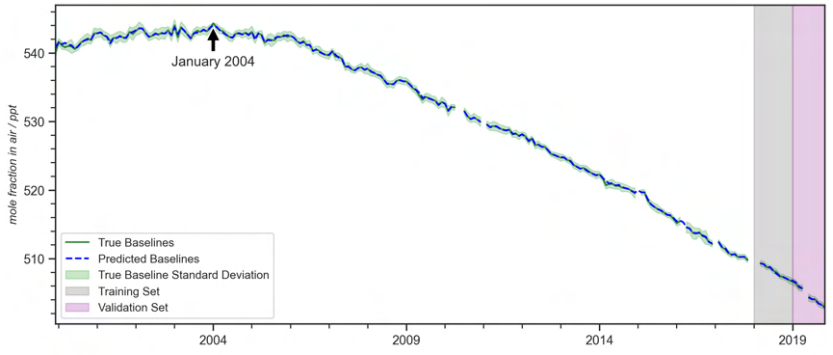


4.10.3 CFC-12

Mole fraction time series

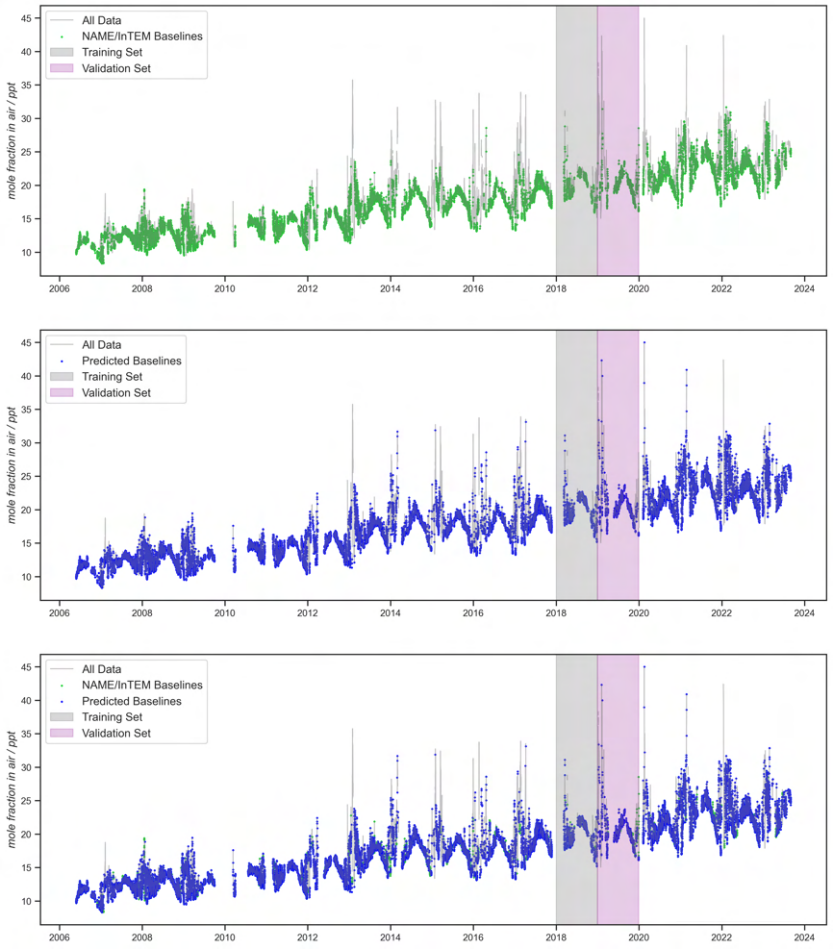


Monthly means

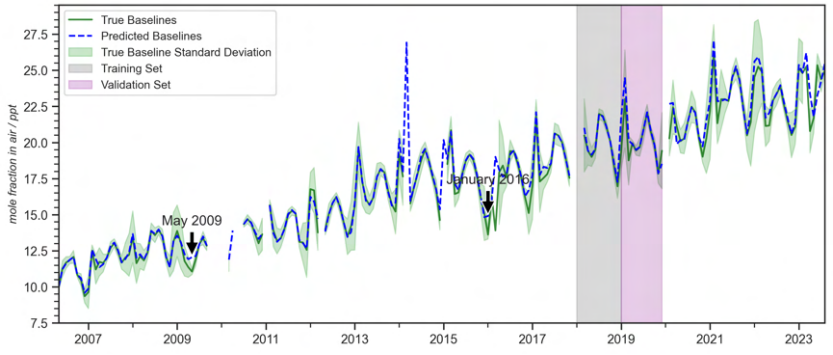


4.10.4 CH_2Cl_2

Mole fraction time series

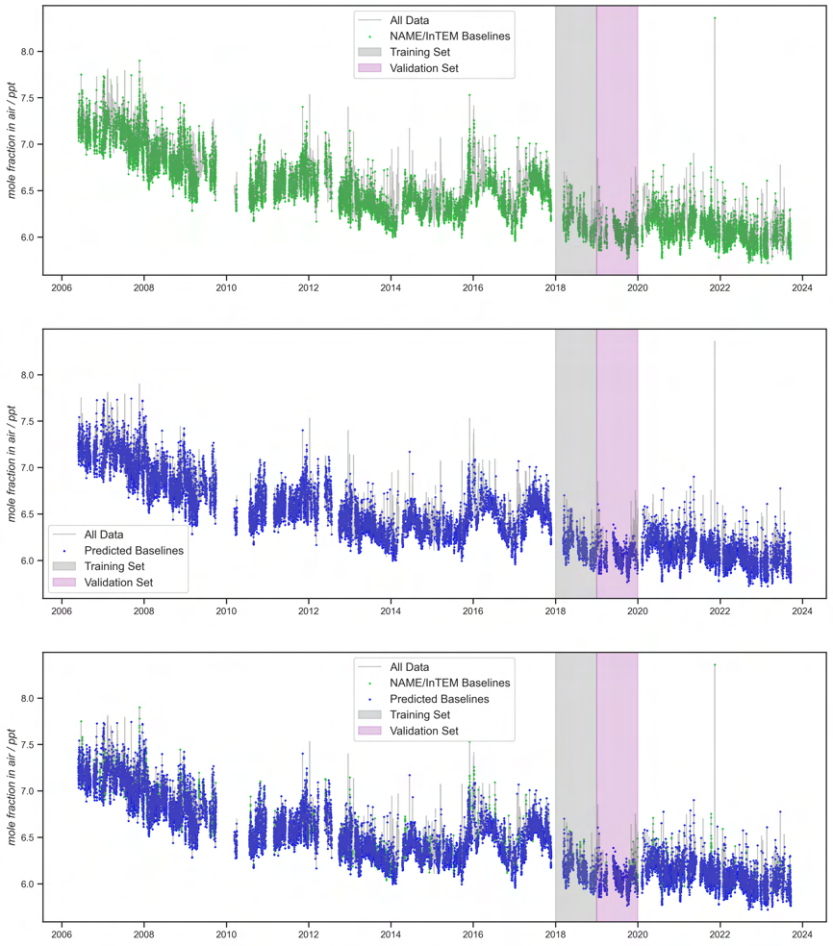


Monthly means

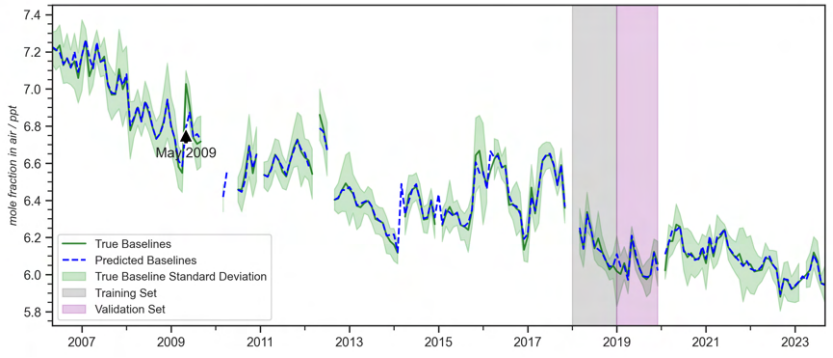


4.10.5 CH_3Br

Mole fraction time series

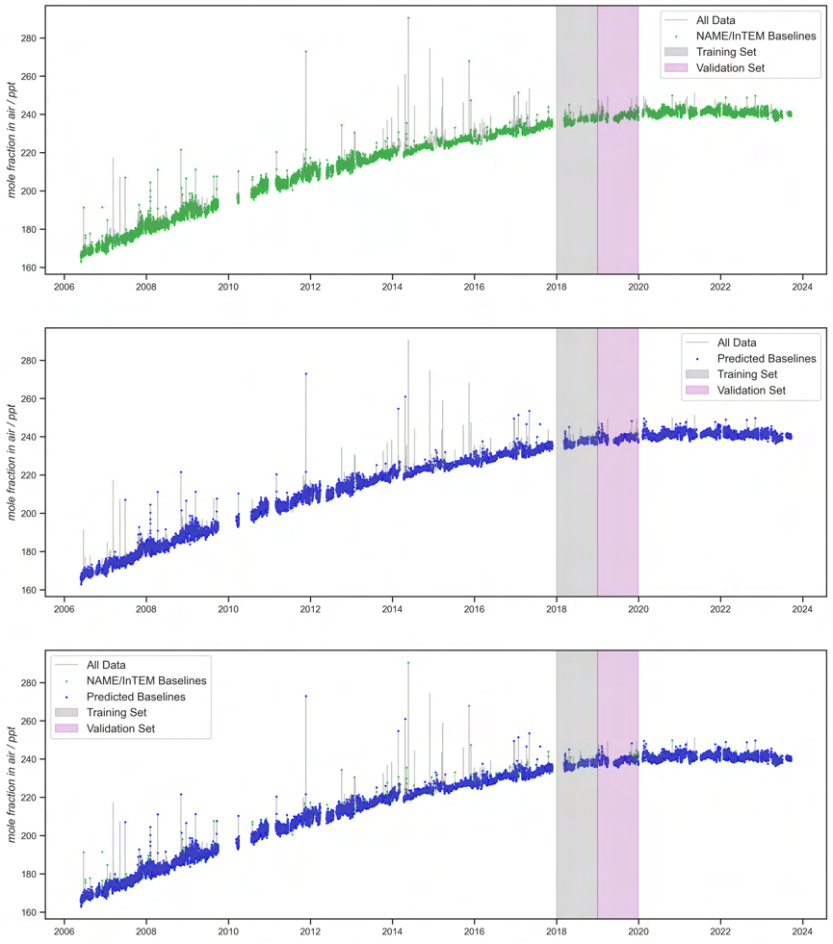


Monthly means

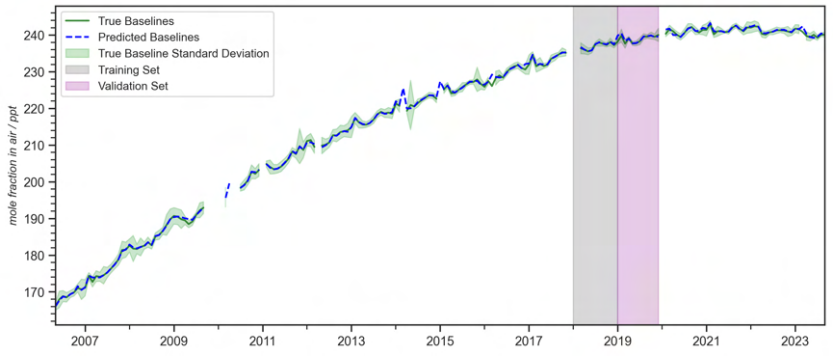


4.10.6 HCFC-22

Mole fraction time series

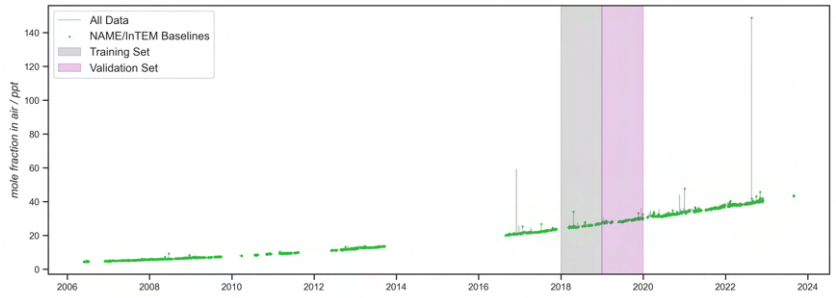


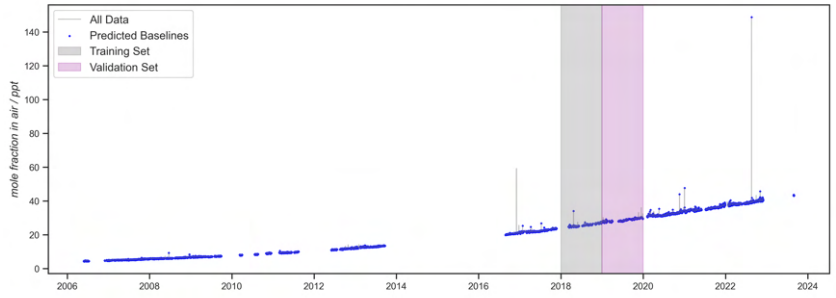
Monthly means

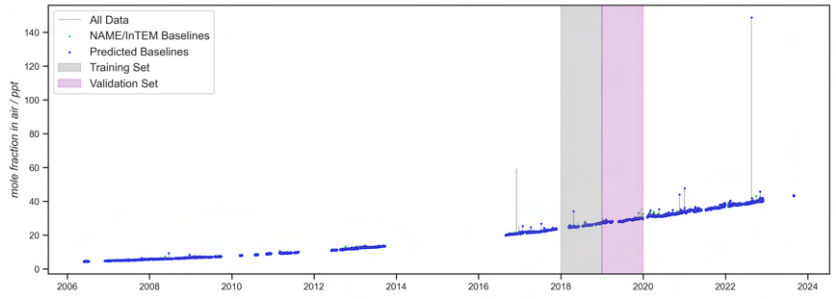


4.10.7 HFC-125

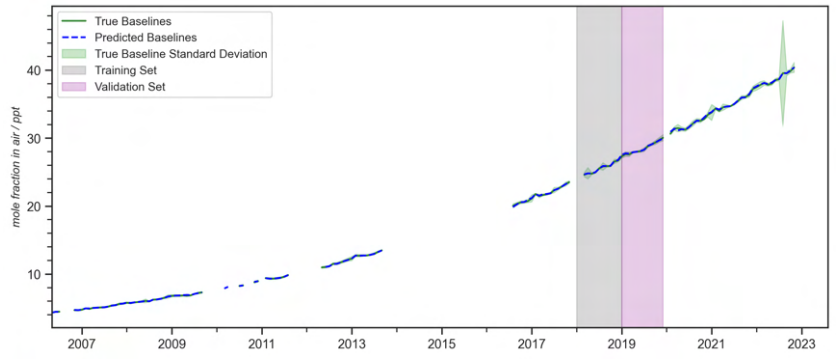
Mole fraction time series





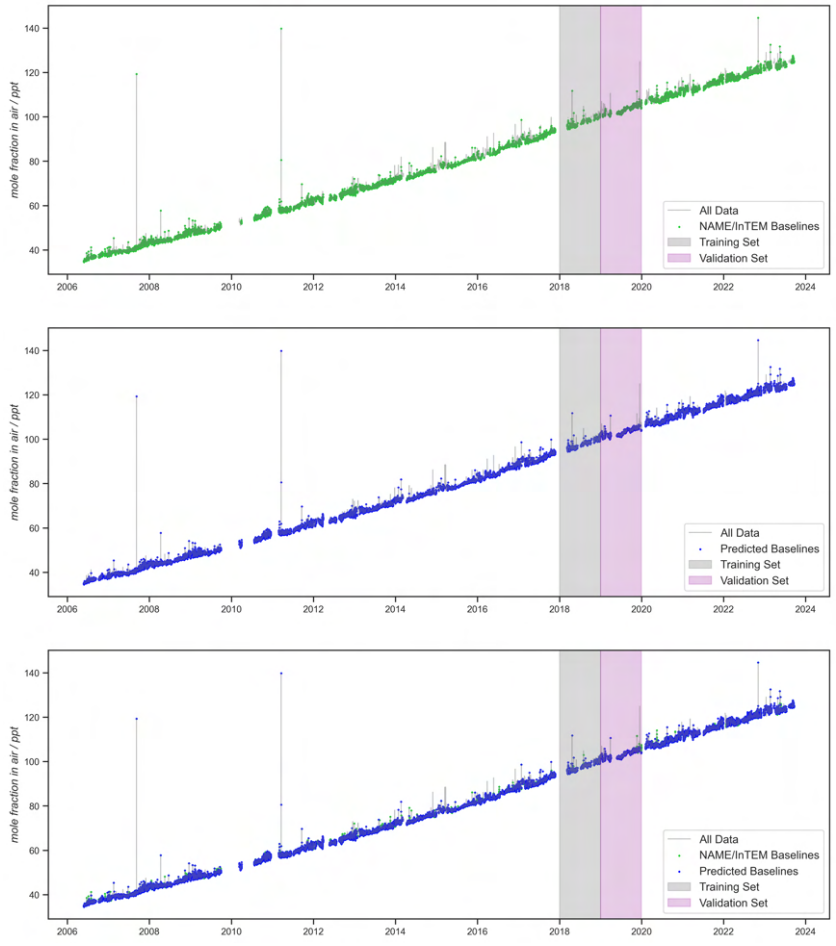


Monthly means

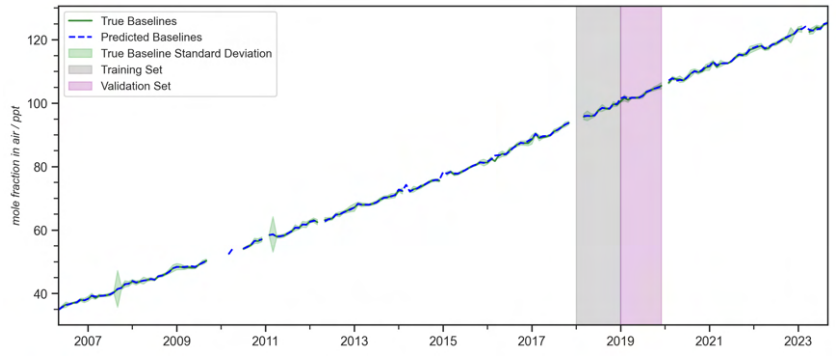


4.10.8 HFC-134a

Mole fraction time series

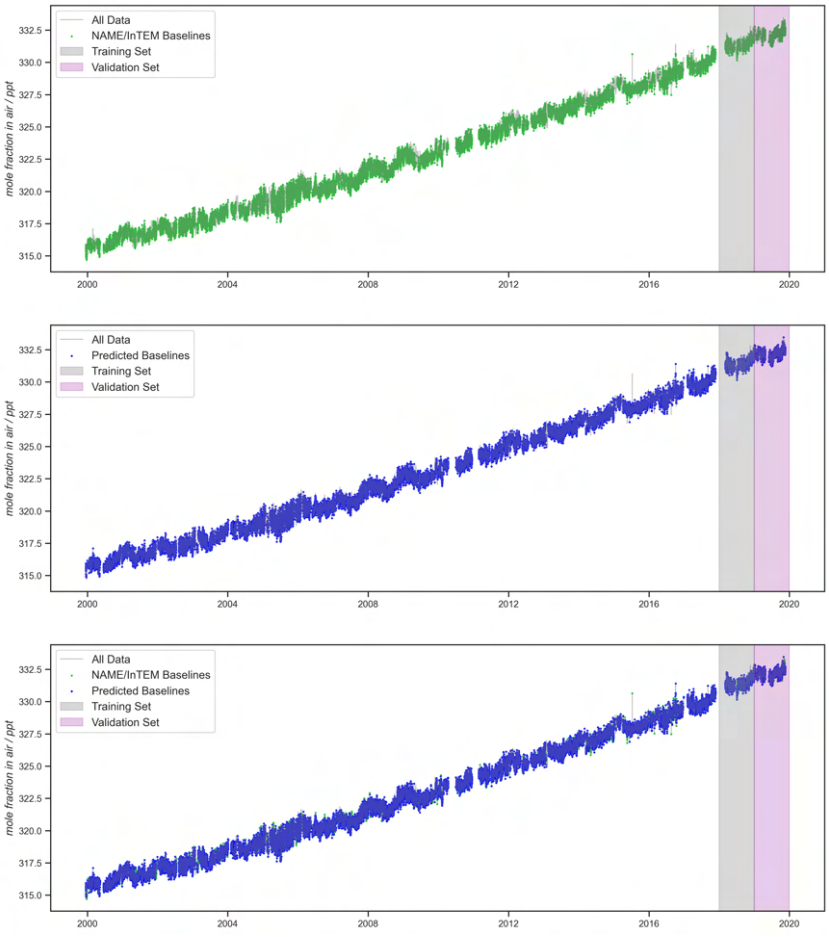


Monthly means

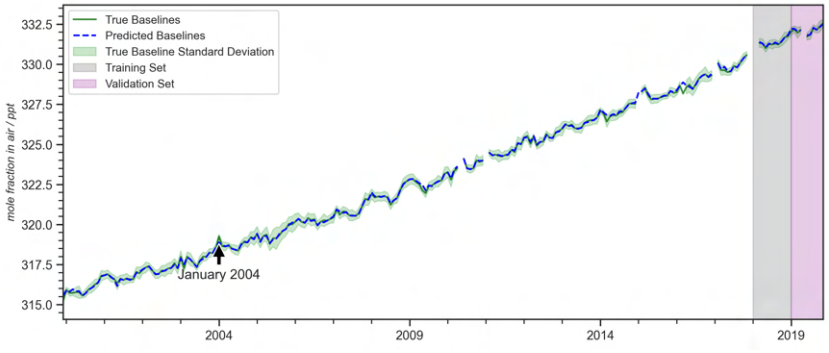


4.10.9 N_2O

Mole fraction time series

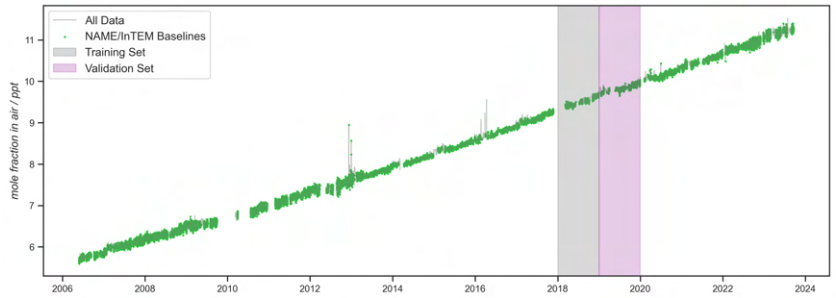


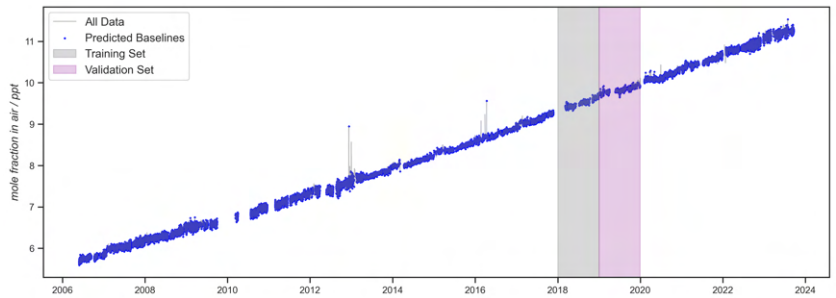
205 **Monthly means**

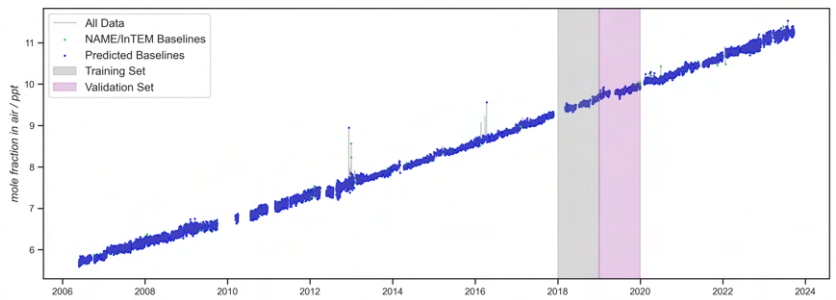


4.10.10 SF_6

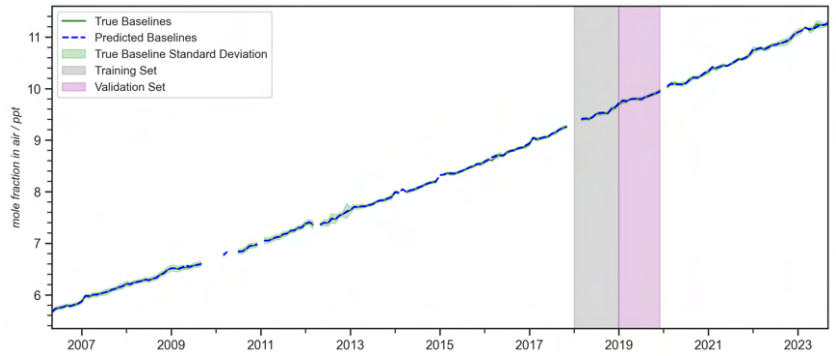
Mole fraction time series







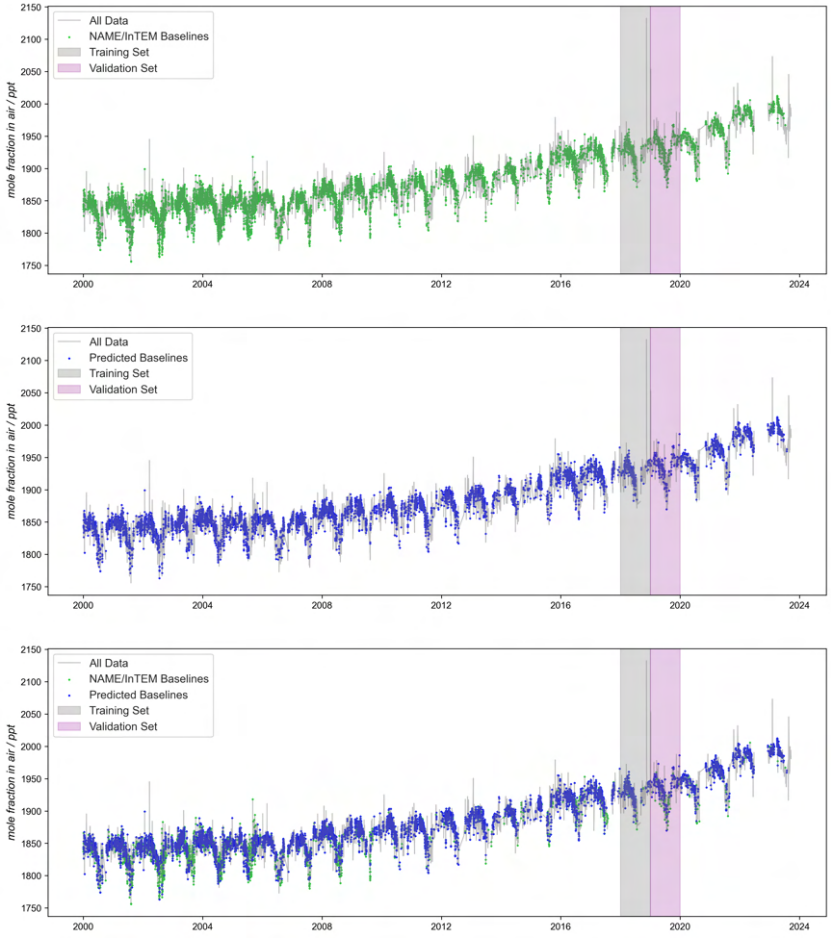
Monthly means



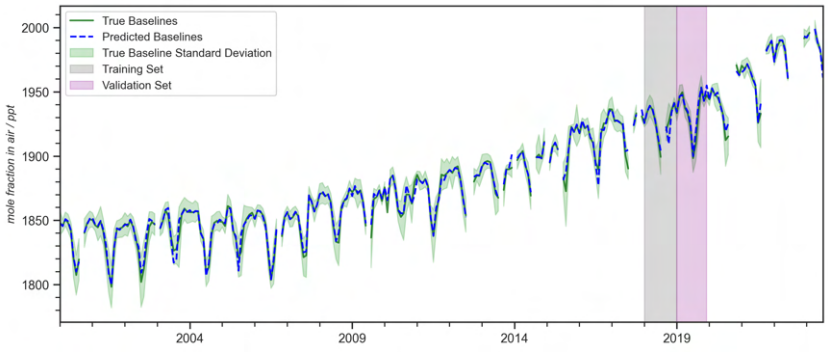
4.11 Trinidad Head, USA

4.11.1 CH_4

Mole fraction time series

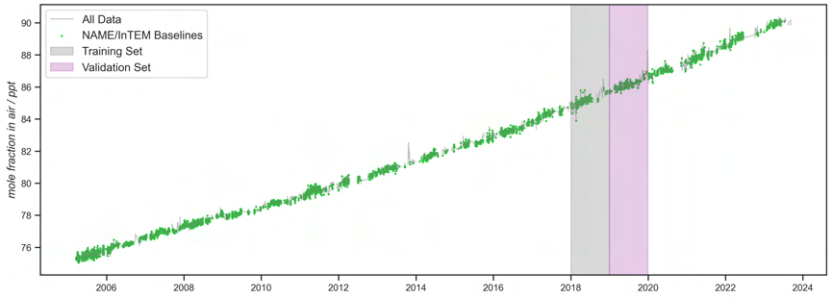


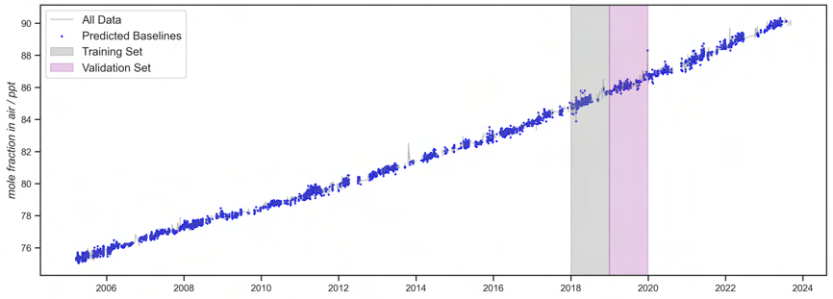
Monthly means

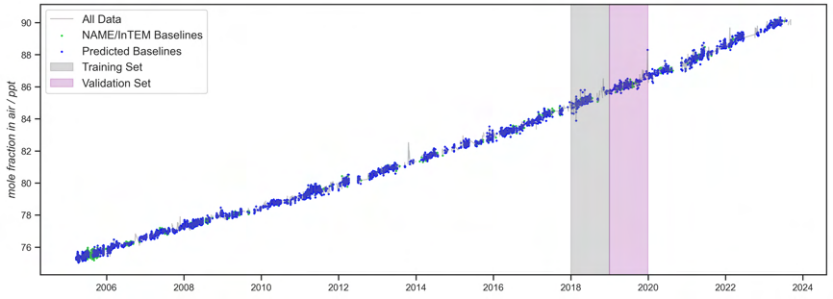


4.11.2 CF_4

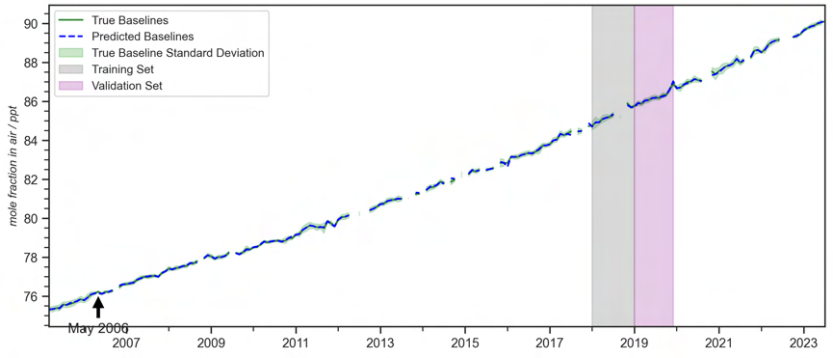
Mole fraction time series





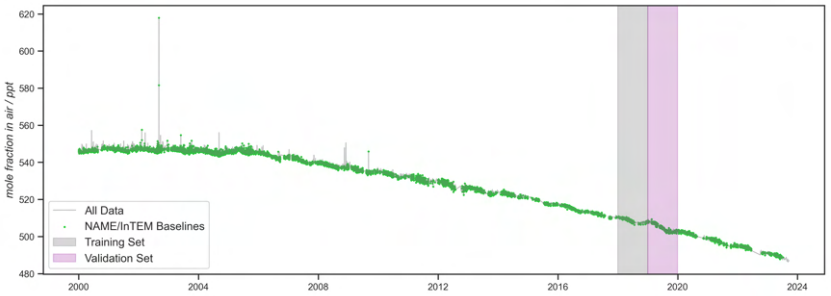


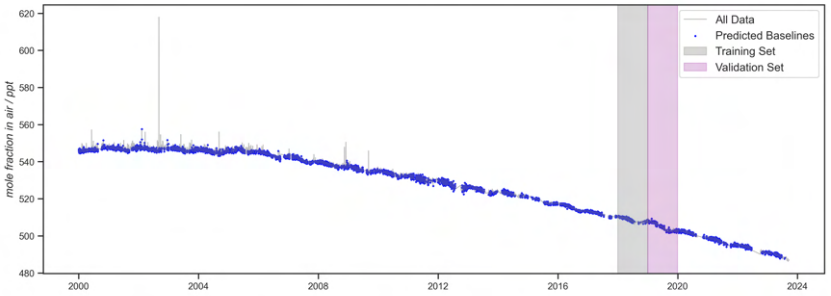
Monthly means

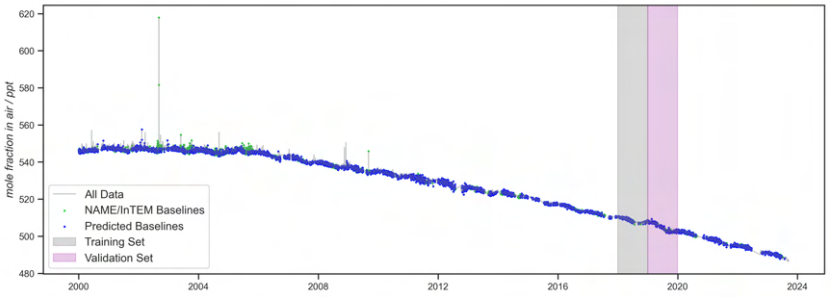


4.11.3 CFC-12

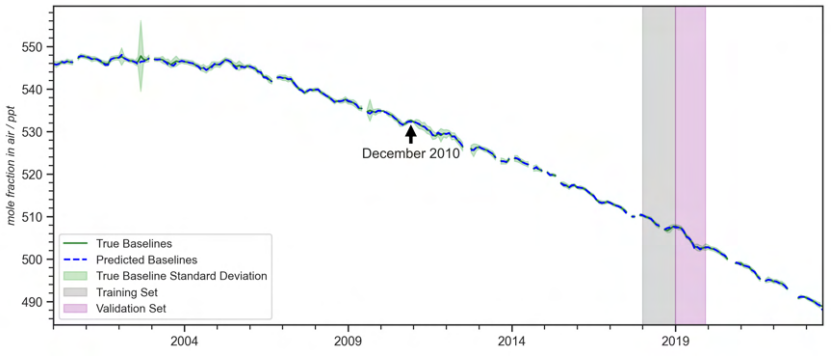
Mole fraction time series





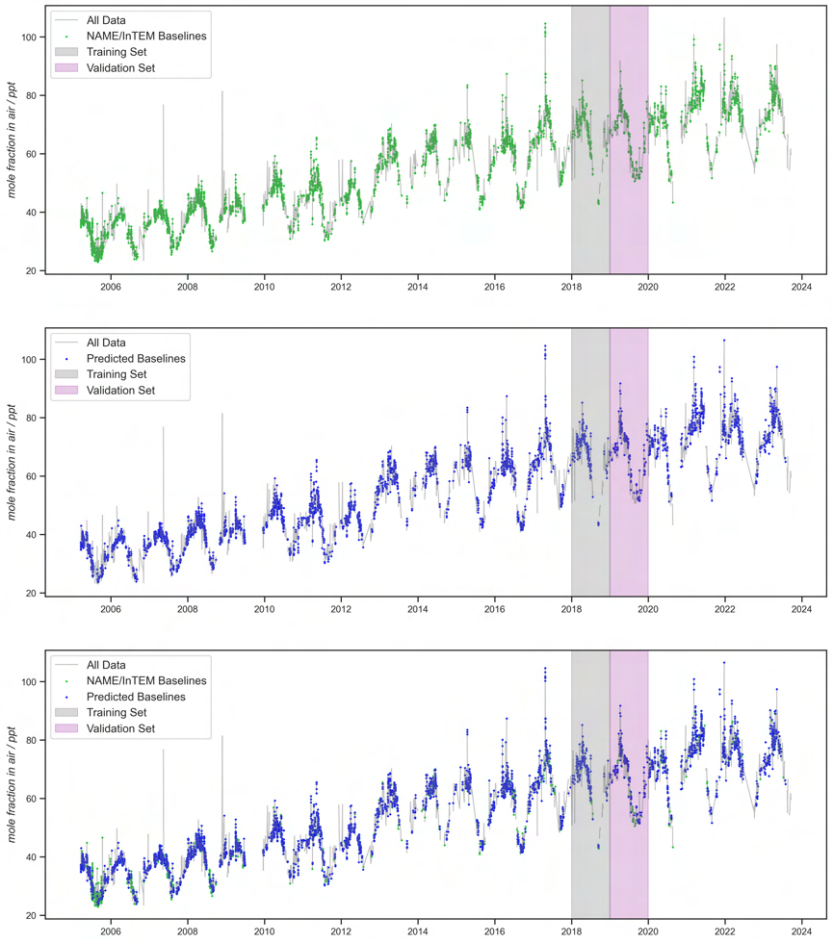


Monthly means

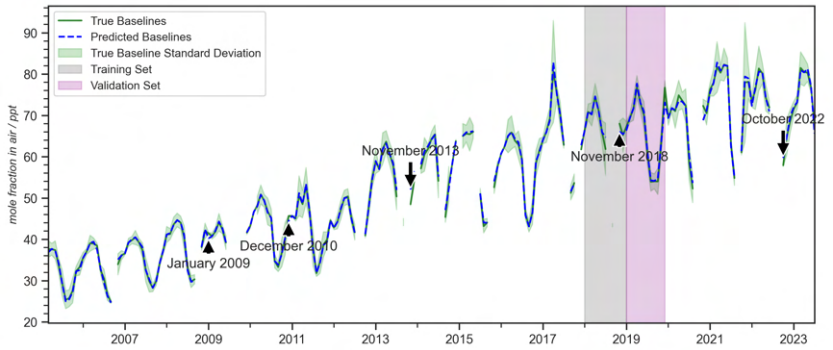


4.11.4 CH_2Cl_2

Mole fraction time series

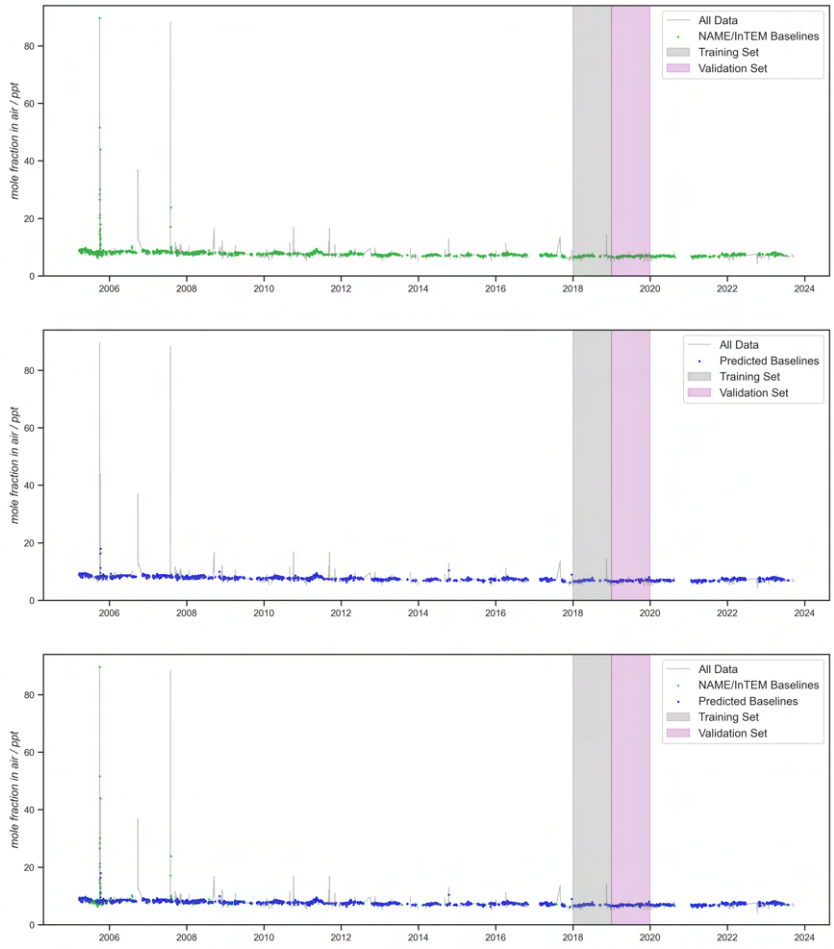


Monthly means

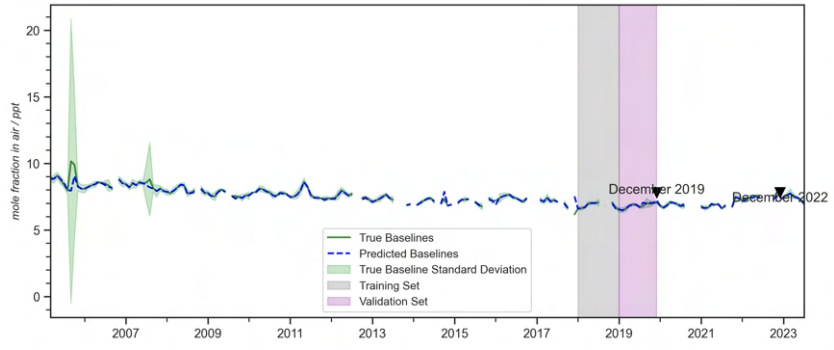


4.11.5 CH_3Br

Mole fraction time series

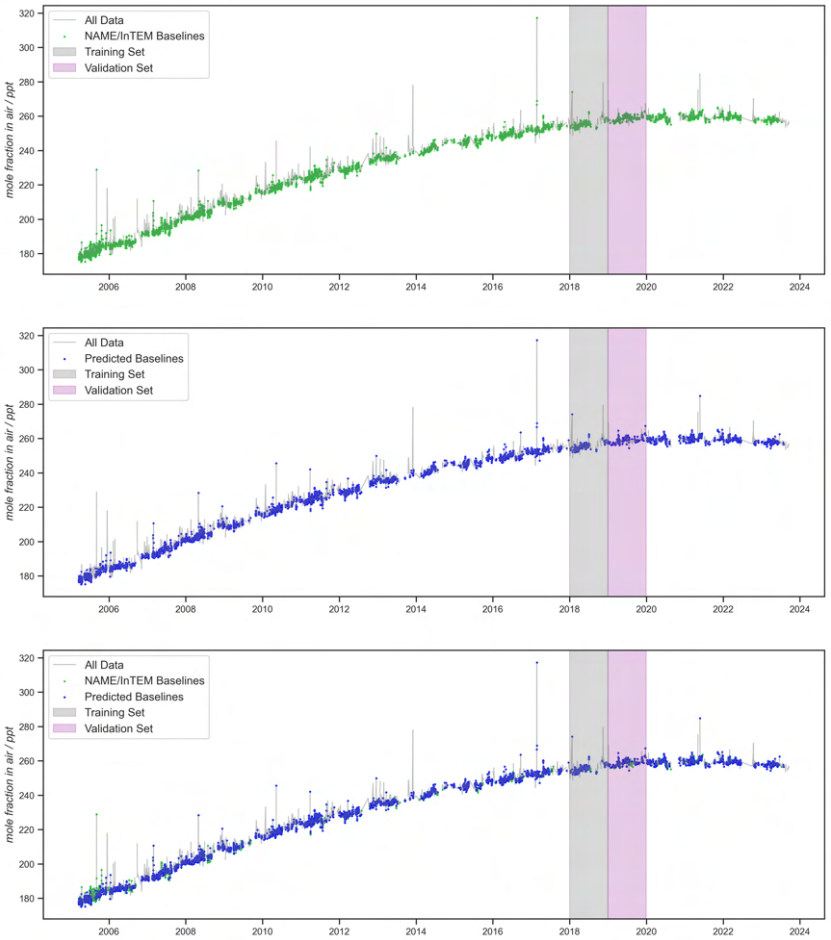


Monthly means

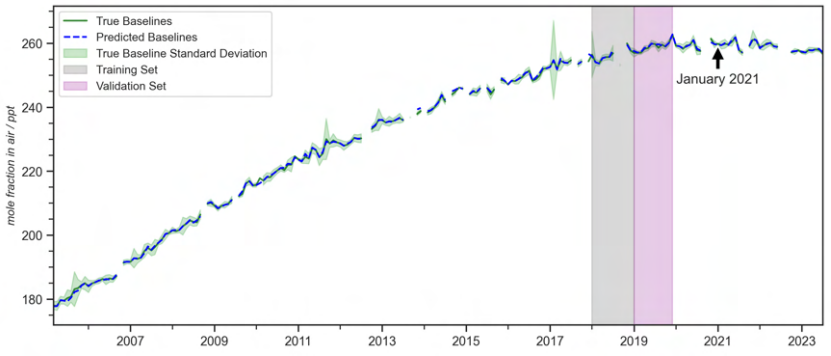


4.11.6 HCFC-22

Mole fraction time series

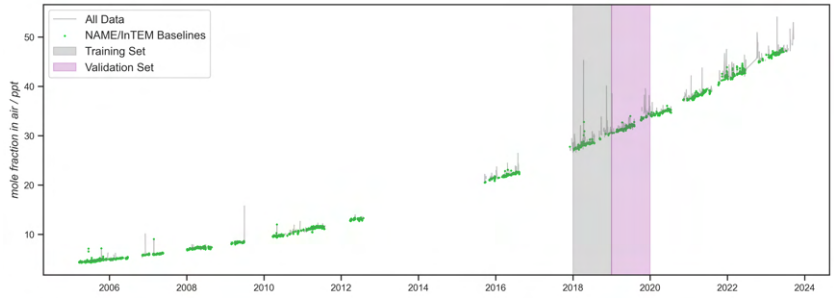


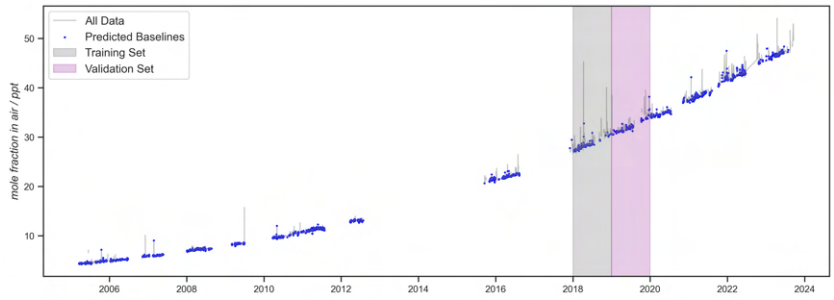
Monthly means

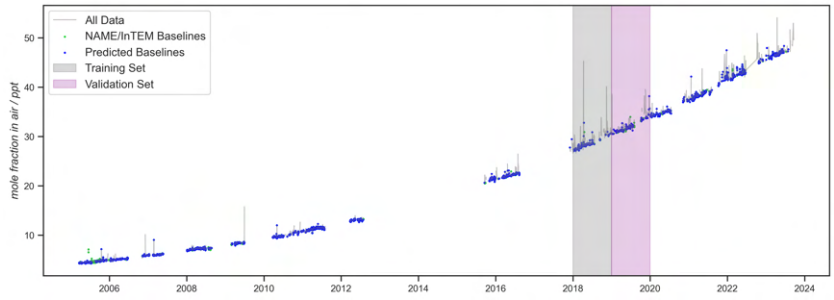


4.11.7 HFC-125

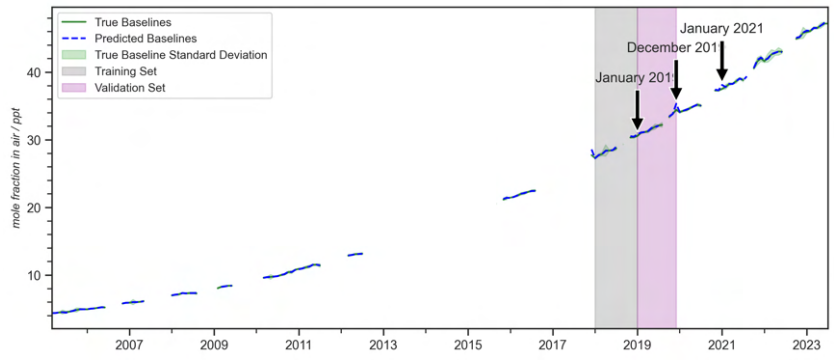
Mole fraction time series





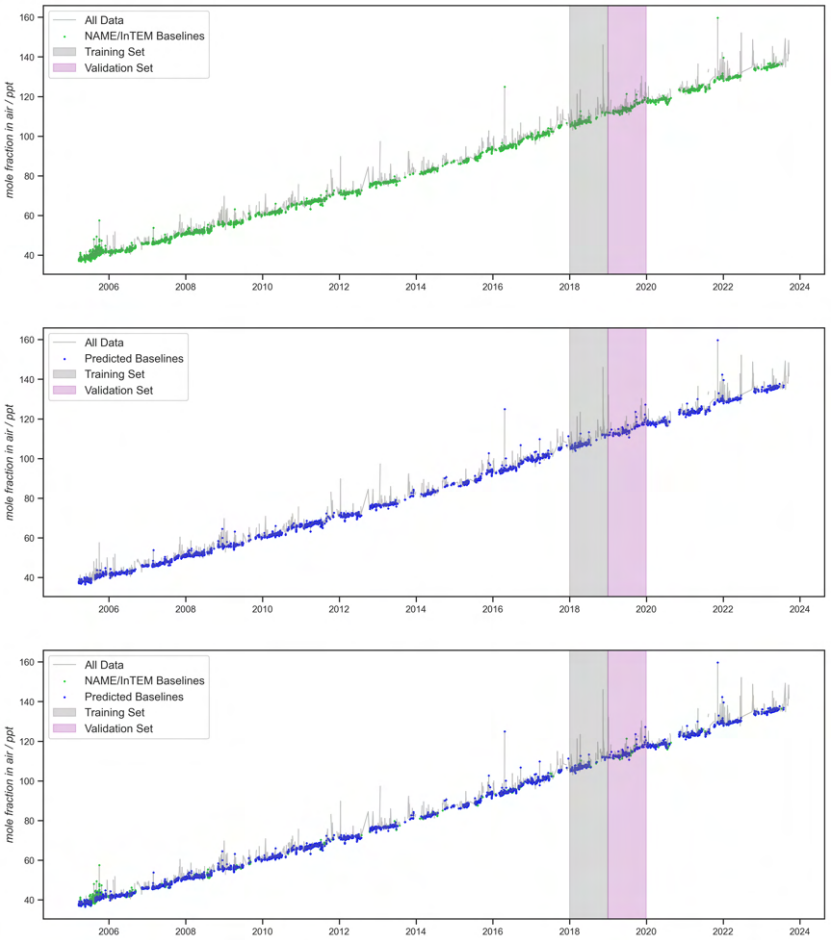


Monthly means

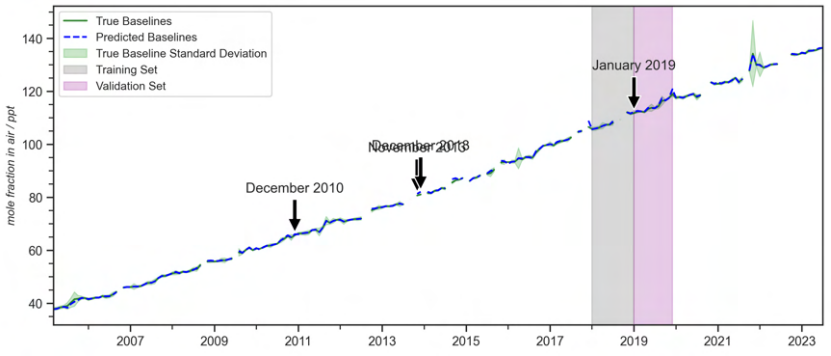


4.11.8 HFC-134a

Mole fraction time series

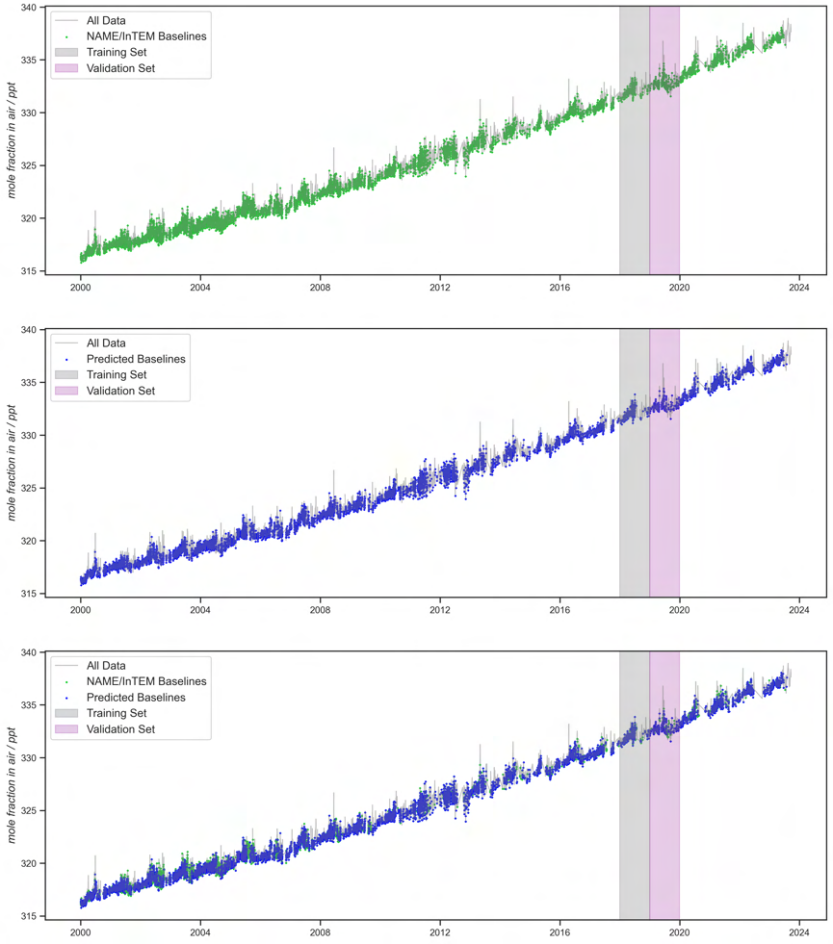


Monthly means

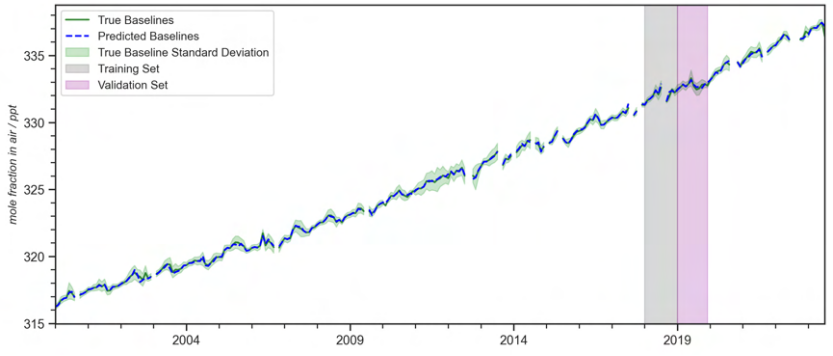


4.11.9 N_2O

Mole fraction time series

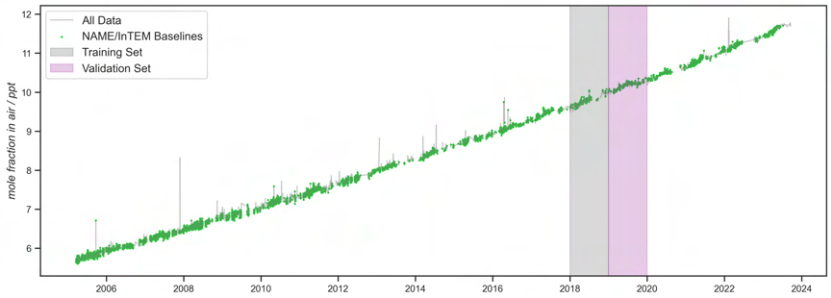


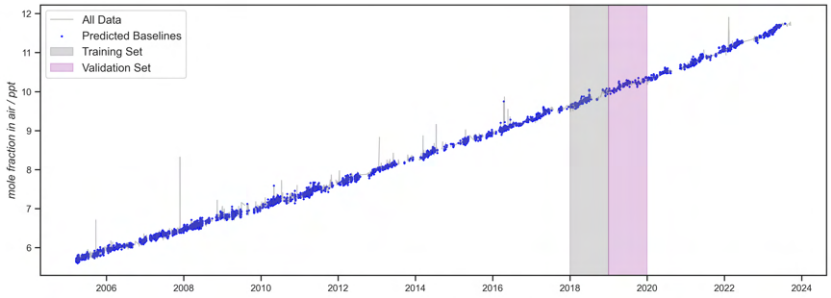
Monthly means

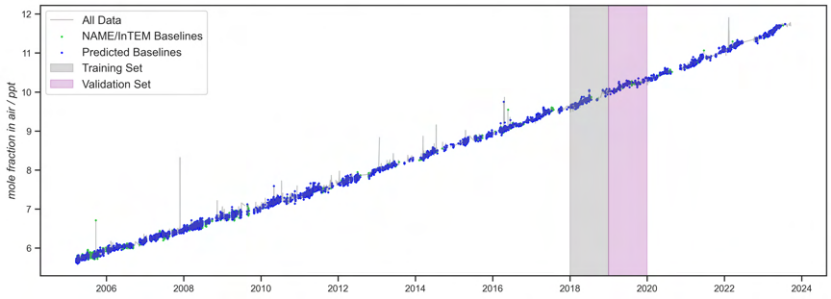


4.11.10 SF_6

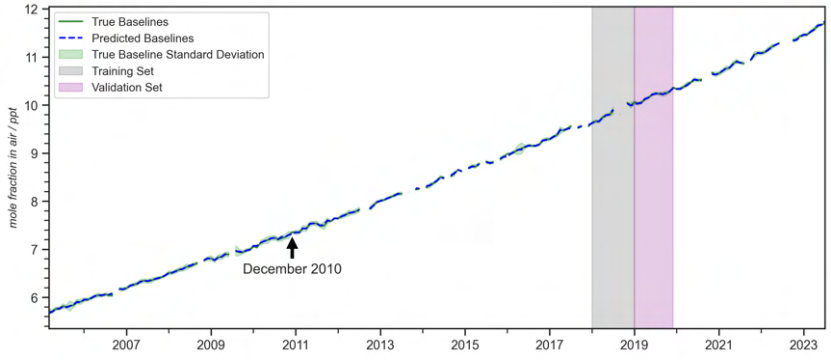
Mole fraction time series







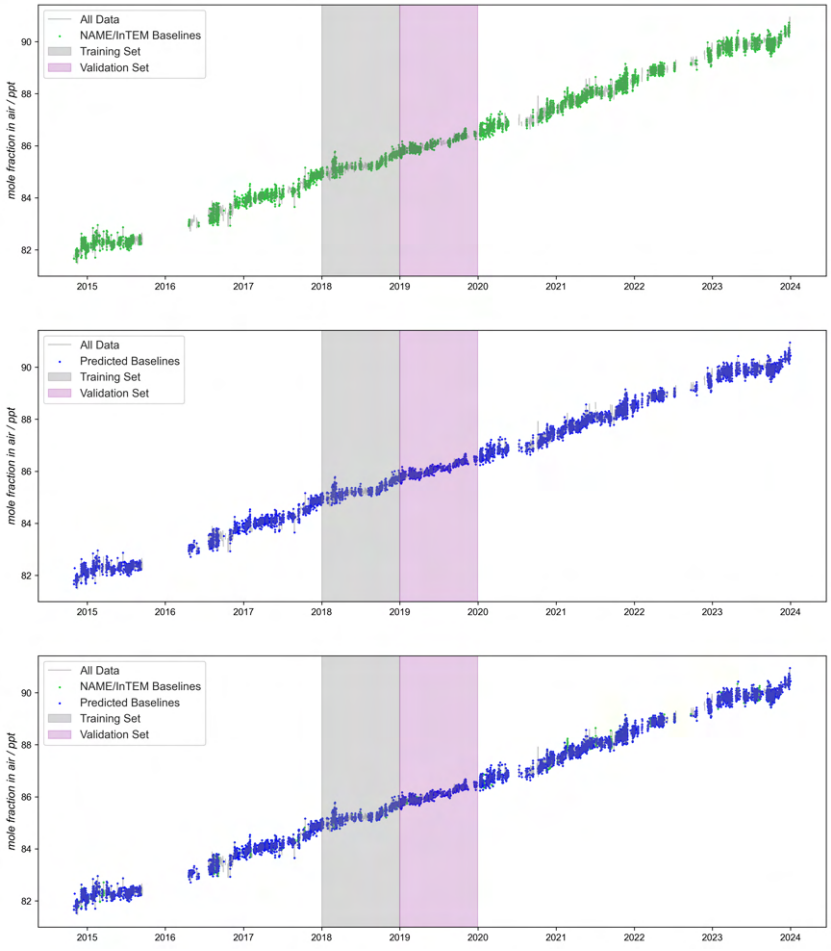
Monthly means



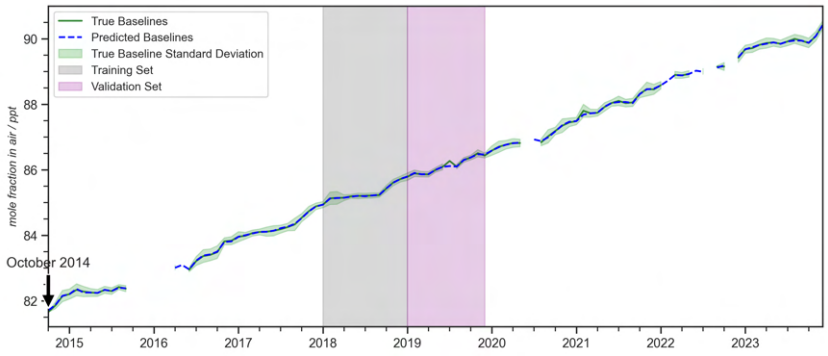
4.12 Zeppelin, Svalbard

4.12.1 CF_4

Mole fraction time series

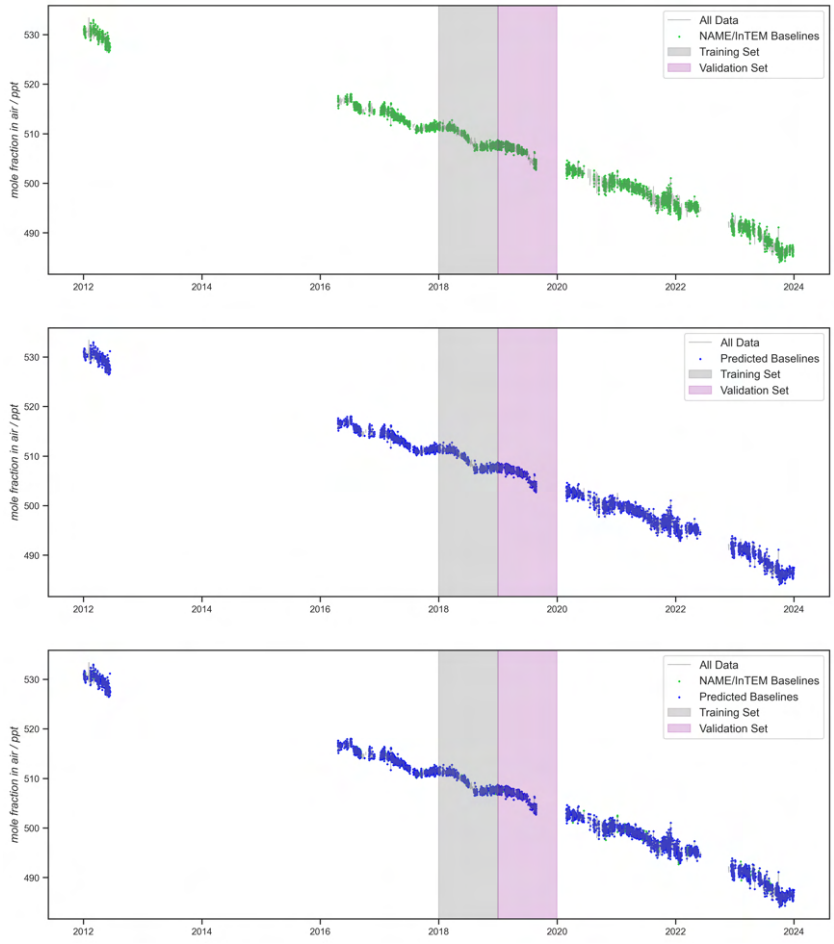


Monthly means

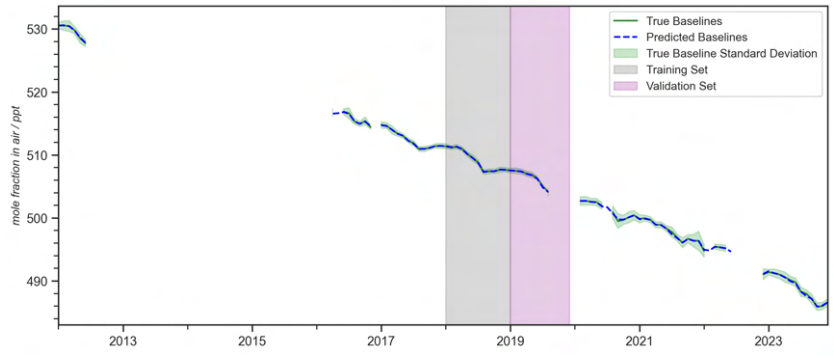


4.12.2 CFC-12

Mole fraction time series

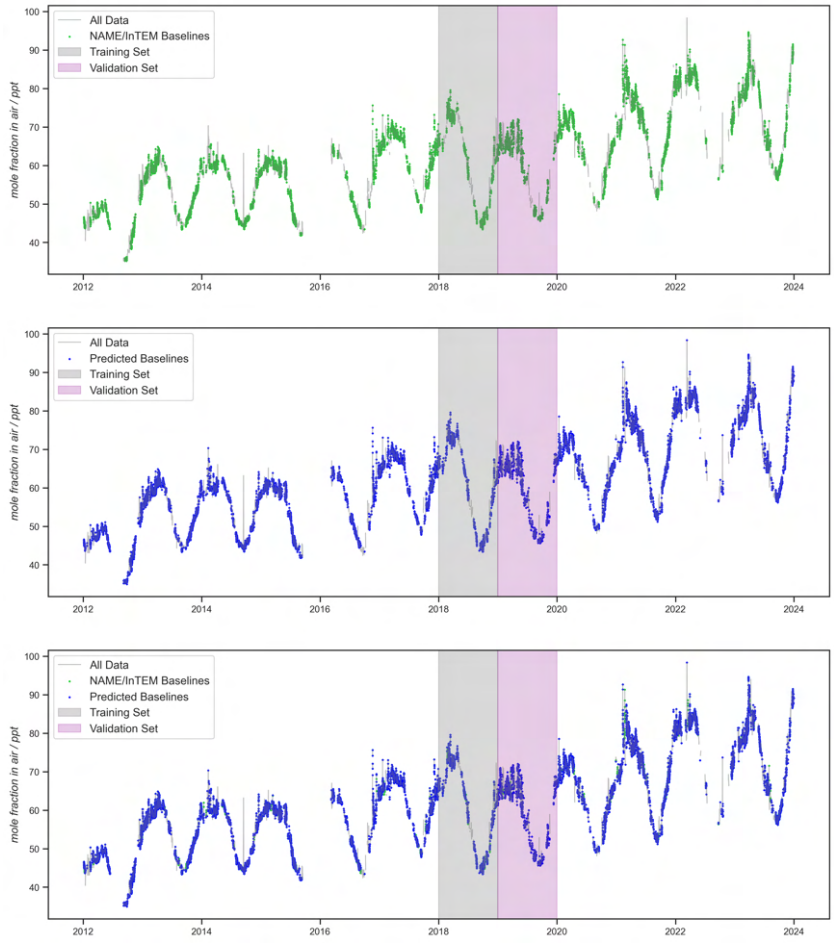


Monthly means

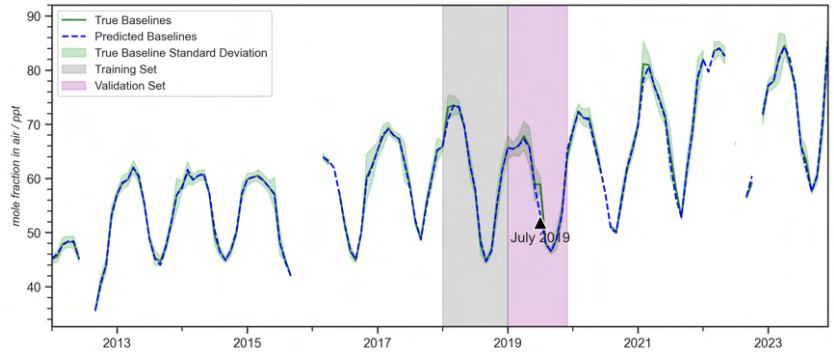


4.12.3 CH_2Cl_2

Mole fraction time series

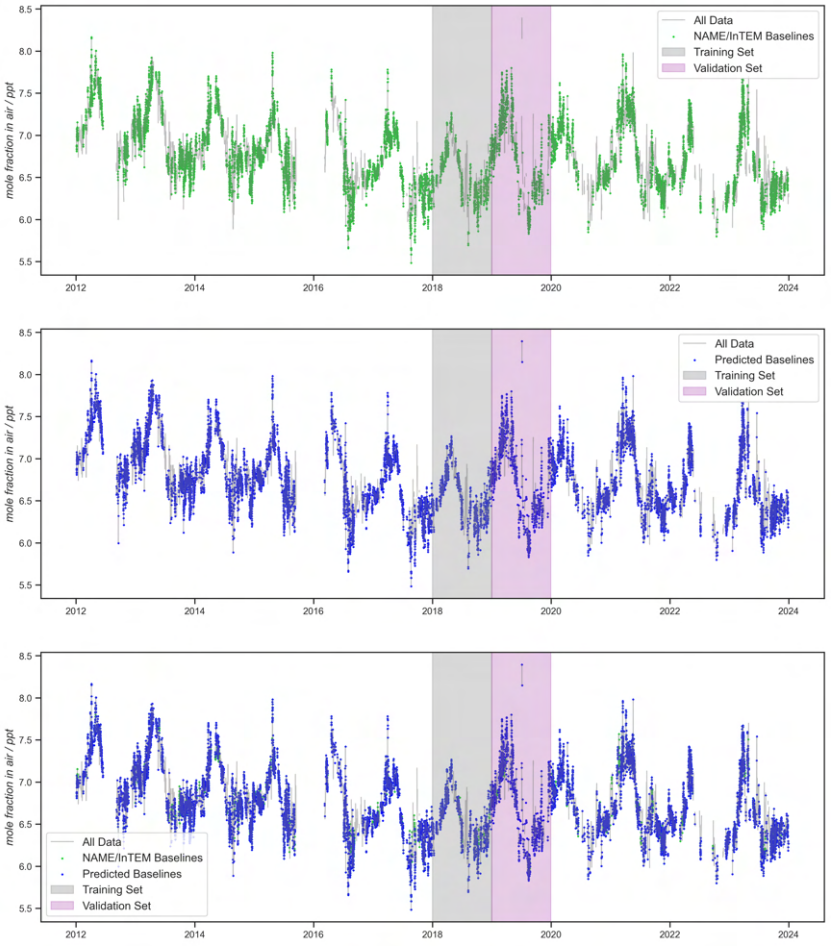


Monthly means

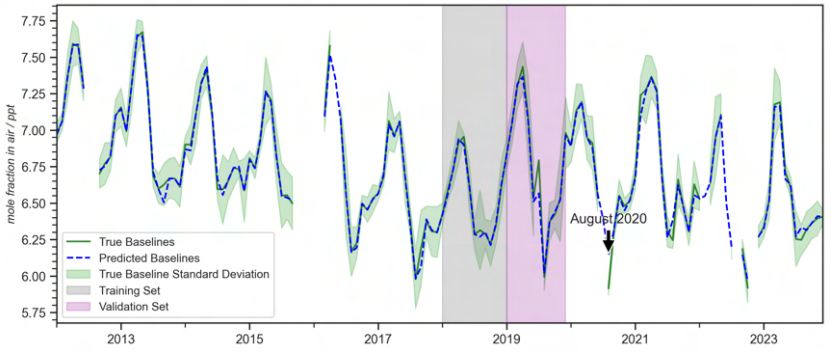


4.12.4 CH_3Br

Mole fraction time series

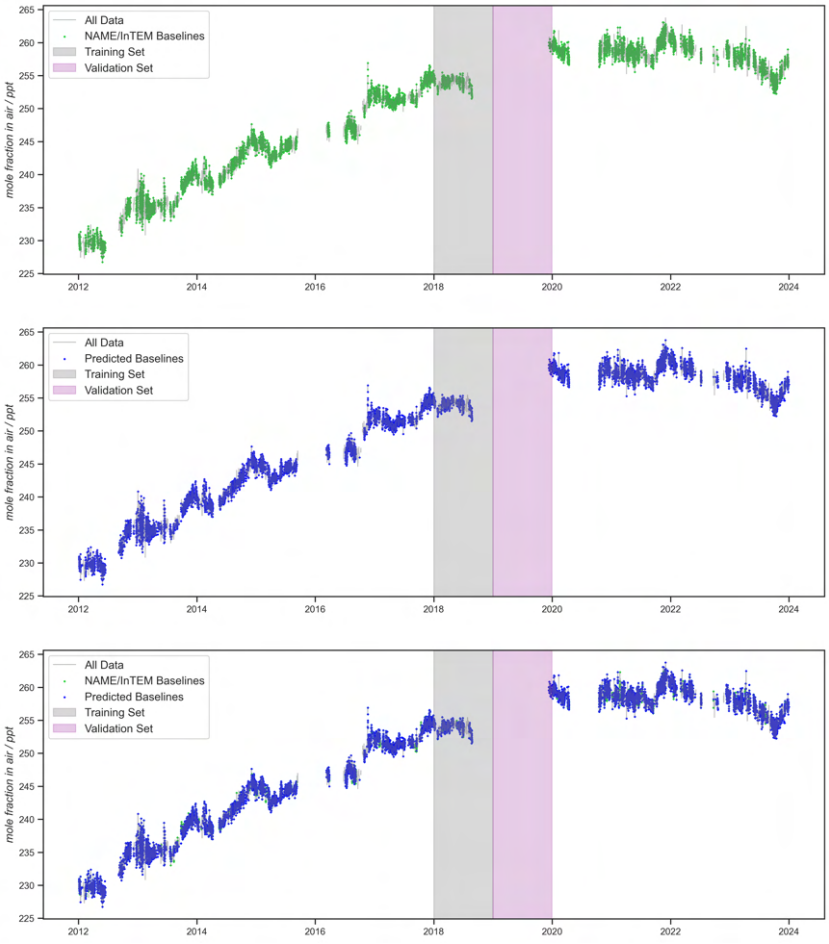


Monthly means

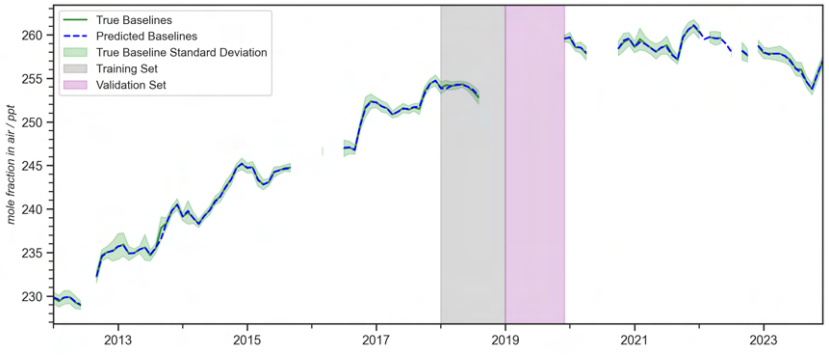


4.12.5 HCFC-22

Mole fraction time series

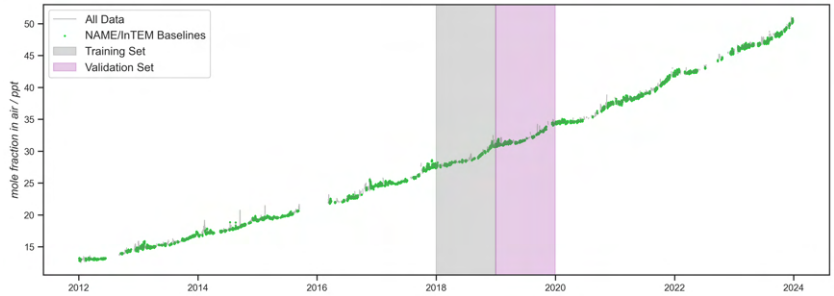


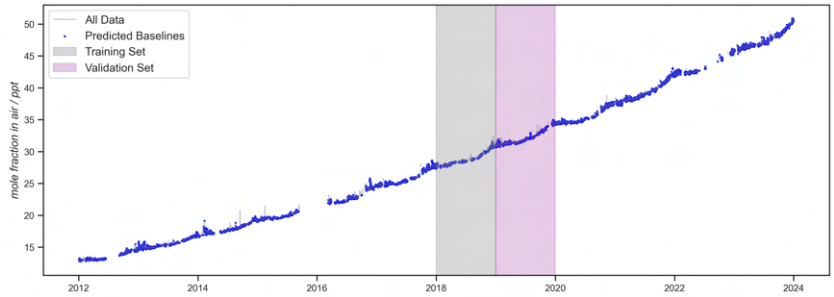
Monthly means

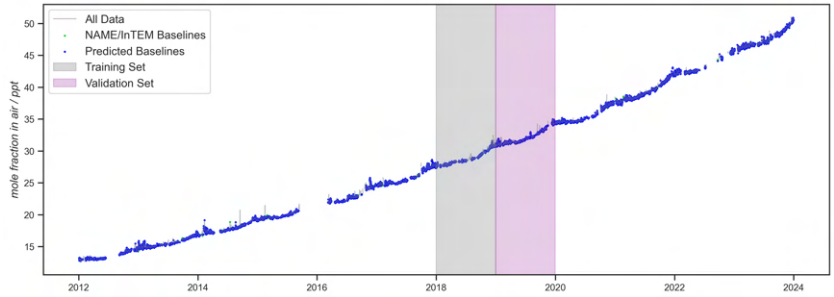


4.12.6 HFC-125

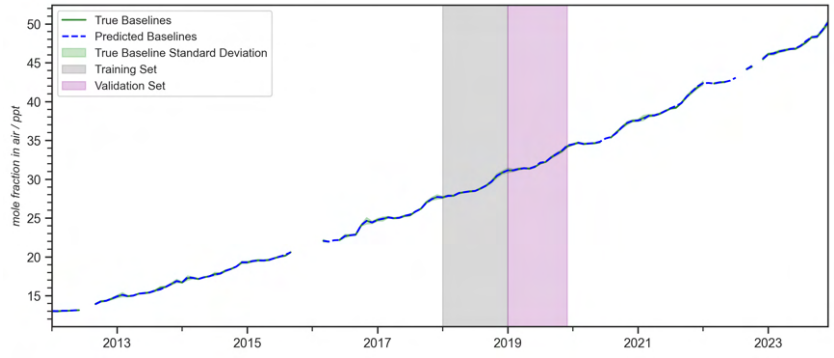
Mole fraction time series





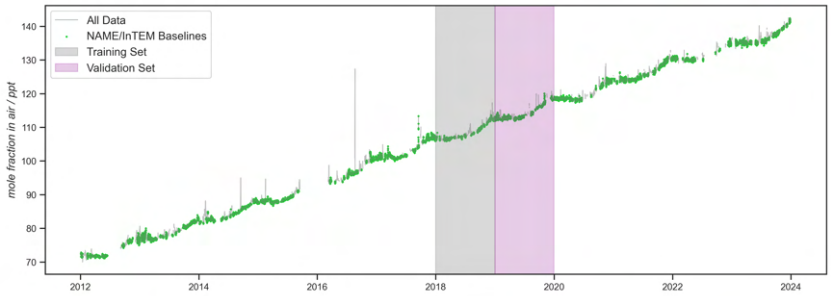


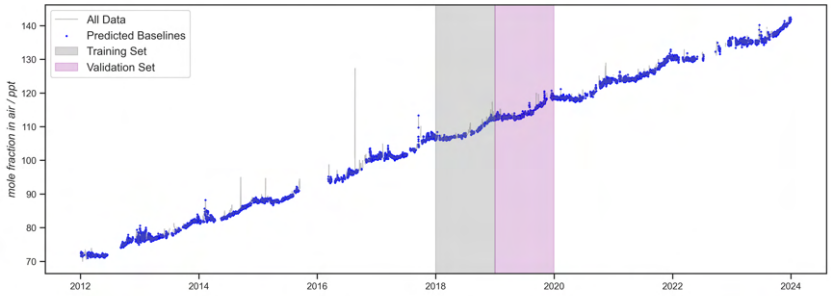
Monthly means

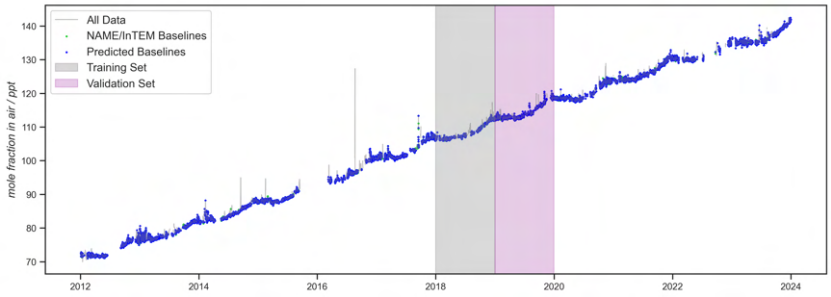


4.12.7 HFC-134a

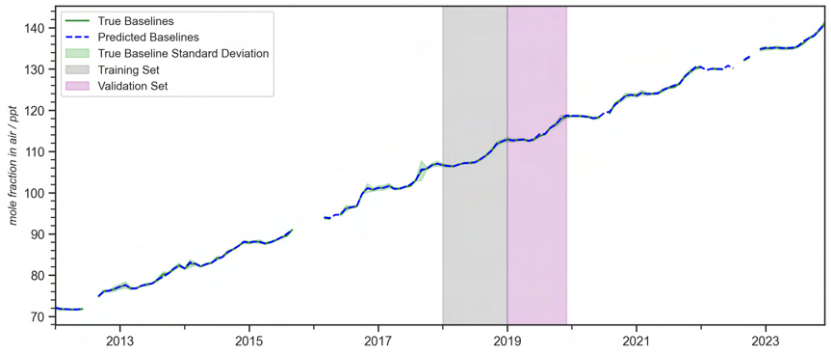
Mole fraction time series





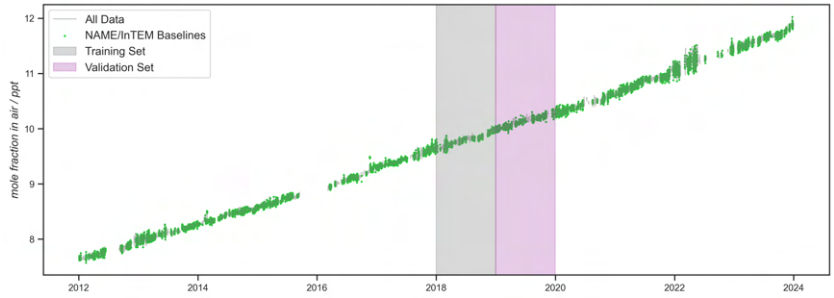


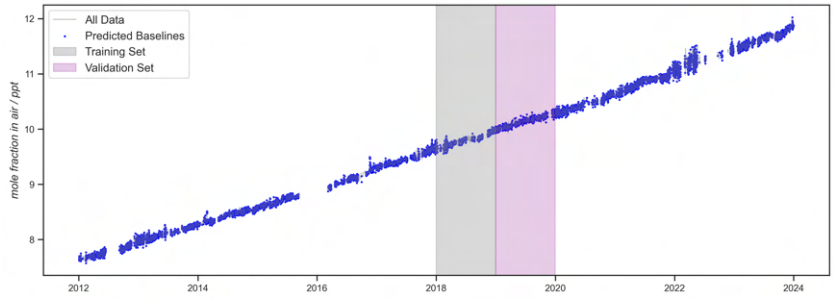
Monthly means

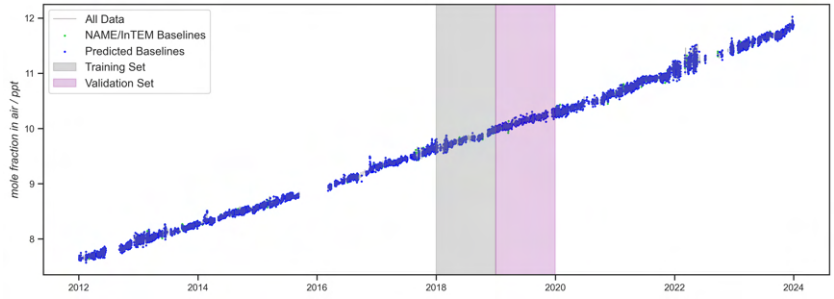


4.12.8 SF_6

Mole fraction time series







Monthly means