Differences between CanOSSEM estimates version 1 and 2

CanOSSEM estimates version 2

Tools such as CanOSSEM require re-training annually, after a major wildfire event or new data becomes available. We re-trained CanOSSEM to specifically address two points:

1. Spatial grid expansion to include remote locations that were missed out in version 1.
2. Time series extension to cover the years 2020, 2021 and 2022.

Spatial grid expansion

We expanded the CanOSSEM spatial grid to cover the locations of remote population centers. It involved the completion of the following steps:

i. Verified intersection of CanOSSEM raster cell with a postal-code single link indicator (SLI).

ii. All 698 indigenous communities across Canada were identified on CanOSSEM population grid.

iii. NASA nightlight data product, VNP46A4 from the black marble suite was used to capture any missed remote population center.

In total, 65,497 new CanOSSEM raster cells were identified and were then used in addition to the 131,974 populated cells for modeling and estimate generation.

Time series extension

CanOSSEM time-series was extended to include the years 2020-2022. Over 63,000 files sized 400+ GB were downloaded and were processed for extraction and transformation. We processed CanOSSEM raster data for each day such as:

• 970,215 CanOSSEM raster cells x 4,748 days (2010-2022) → 4,606,580,820 data rows for 30 variables

• After filtering for populated raster cells only: ~ 936 million data rows

Dataset size difference between version 1 and 2

CanOSSEM version 2 was trained using almost 3 times more training data.

Number of dataset rows increased 3 times for CanOSSEM version 2 compared to version 1.

Datasets	CanOSSEM version 1	CanOSSEM version 2
Complete Dataset ($C_{100\%}$)	17,890,113	44,321,329
Prediction Set ($P_{10\%}$)	1,789,011	4,432,132
Reduced Dataset ($R_{54\%}$)	9,660,661	23,931,012
Validation Set ($V_{11\%}$)	1,932,132	4,786,202
Training Set ($T_{43\%}$)	7,728,529	19,144,810

Performance metrics difference between version 1 and 2

RandomForest algorithm was used for model training through ranger package.

Important model nomenclature:

- M1: Primary CanOSSEM version 2 with expanded grid and extended time-series.
- M2: Secondary CanOSSEM version 2 (trained without AOD variables) with expanded grid and extended time-series.

M1 performance compared against Primary CanOSSEM (version 1). About 18% improvement in the RMSE was calculated for the M1.

Minor improvement in M2 performance was seen when compared against Secondary CanOSSEM (version 1).

Correlation between observed and predicted $PM_{2.5}$

High correlation seen in $P_{10\%}$ dataset between observed and predicted $PM_{2.5}$ for Primary CanOSSEM version 2 (M1)

Summary statistics for estimates by year

Year	Range	Variance	Standard Deviation	Interquartile Range	Coefficient of Variation
2010 - v1	249.45	12.47	3.53	2.62	0.48
2010 - v2	322.44	13.21	3.63	3.15	0.50
2013 - v1	203.51	7.18	2.68	2.32	0.37
2013 - v2	233.30	8.21	2.86	2.82	0.41
2012 - v1	219.40	20.64	4.54	2.70	0.58
2012 - v2	214.07	18.78	4.33	3.30	0.57
2013 - v1	170.33	8.65	2.94	2.46	0.40
2013 - v2	167.91	12.16	3.49	3.05	0.48
2014 - v1	251.51	9.82	3.13	2.41	0.43
2014 - v2	315.38	14.40	3.79	3.05	0.52
2015 - v1	368.54	37.24	6.10	2.64	0.77
2015 - v2	386.38	33.09	5.75	3.24	0.74
2016 - v1	1229.07	14.00	3.74	2.31	0.53
2016 - v2	1257.85	13.82	3.72	2.86	0.53
2017 - v1	339.91	38.37	6.19	2.54	0.78
2017 - v2	340.15	33.04	5.75	3.17	0.74
2018 - v1	806.25	58.91	7.67	2.80	0.93
2018 - v2	806.42	45.65	6.76	3.42	0.85

Year	Range	Variance	Standard Deviation	Interquartile Range	Coefficient of Variation
2019 - v1	339.11	12.72	3.57	2.40	0.49
2019 - v2	311.63	14.83	3.85	2.96	0.53
2020 - v2	214.86	8.96	3.00	2.27	0.43
2021 - v2	455.33	32.13	5.67	2.66	0.71
2022 - v2	181.27	10.01	3.16	2.47	0.43