DYNAMIC NOWCAST OF
THE NEW ZEALAND GREENHOUSE GAS INVENTORY

MALCOLM JONES, HANNAH CHORLEY, FLYNN OWEN, TAMSYN HILDER, HOLLY TROWLAND,
AND PAUL BRACEWELL

ABSTRACT. As efforts to mitigate the effects of climate change grow, reliable and thorough reporting of greenhouse gas emissions are essential for measuring progress towards international and domestic emissions reductions targets. New Zealand's national emissions inventories are currently reported between 15 to 27 months out-of-date. We present a machine learning approach to nowcast (dynamically estimate) national greenhouse gas emissions in New Zealand in advance of the national emissions inventory's release, with just a two month latency due to current data availability. Key findings include an estimated 0.2% decrease in national gross emissions since 2020 (as at July 2022). Our study highlights the predictive power of a dynamic view of emissions intensive activities. This methodology is a proof of concept that a machine learning approach can make sub-annual estimates of national greenhouse gas emissions by sector with a relatively low error that could be of value for policy makers.

1. INTRODUCTION

Emissions from human activity are the key driver of the anthropogenic climate change observed since the onset of the industrial revolution (c. 1850). In an effort to mitigate our influence over the global climate, the Paris Climate Agreement aims to limit global temperature rise to well below 2°C and pursue efforts to limit warming to 1.5°C above pre-industrial levels, requiring significant action to reduce greenhouse gas emissions to net zero on an international scale. Reliable and thorough greenhouse gas inventories are fundamental in measuring progress towards these targets and provide insight into domestic and international emissions reduction policies and strategies [53]. Guidelines for such inventories are published by the Intergovernmental Panel on Climate Change (IPCC) and were refined in 2019, the first major update since 2006 [15]. We develop a proof of concept that combines machine learning methods with indicators of dominant trends in New Zealand's emissions profile in order to improve timeliness of emissions reporting with relatively low error, and therefore 'nowcast' New Zealand's emissions in the sense described in [26]. The importance of greater timeliness is also highlighted in [53].

New Zealand's national emissions are reported in the New Zealand Greenhouse Gas Inventory (NZGGI). It is the official annual report of all anthropogenic emissions and removals of greenhouse gases (GHGs) associated with production, which is currently released approximately 15 months after the end of the calendar year being reported on, as determined by international reporting guidelines [34]. Multiple GHG types estimated by the NZGGI are aggregated in terms of CO_2 equivalents (CO_2-e).[1] Our indicator-based model estimates CO_2-e directly rather than modelling each GHG type to compensate for limited data availability.

National emissions inventories are time-consuming to produce. At any given time, national GHG emissions reported in the NZGGI are between 15 and 27 months out of date. Methodologies for the release of quarterly and more timely emissions estimates are in the process of being developed in a number of countries [24, 26, 46]. Within New Zealand, Statistics New Zealand are developing quarterly emissions reporting [47] but, as we discuss in Section 4.1.4, this is not comparable to the NZGGI in many ways. This existing gap highlights the need for a more dynamic dataset that complements the NZGGI to be developed.

Numerous studies exist that estimate daily emissions (i.e. emissions occurring within any given day) by first developing a daily inventory of emissions. [20] develops a 'bottom-up' inventory using

[1]Various research debates ways of aggregating GHGs (see [44] and references therein). We are bound to CO_2-e since it is used by the NZGGI.

national mobility data, finding that emissions reduced during the COVID-19 pandemic, by as much as 30% during April 2020. [31] constructs a measure of CO_2 emissions by developing an inventory by country comprised of the sectors that are power generation, industry, transportation and household consumption, each broken down by the three fossil fuel types which are coal, oil and natural gas. Findings include that CO_2 emissions decreased by 8.8% in the first half of 2020 compared to the same period of the previous year in China. [43] again develops an inventory using a 'bottom-up' approach, comprised of six sectors, finding that daily CO_2 emissions decreased by -17% by April 2020 compared to the mean 2019 levels, with surface transport being the sector that influenced reduction most heavily. Further studies that estimate daily emissions by developing a daily inventory of emissions include [30, 39, 54]. Such inventories may be labour-intensive to curate, even more so than the NZGGI at the annual level. These studies have impressively specific results, such as daily emissions estimates with a regional component while separating CO_2 and air pollutants [24]. However, they need to focus on short windows of time to obtain a sufficiently detailed inventory in the first place. This distinguishes our work, which reports on emissions from 1990 up to two months prior to the present day.

We produce daily estimates of cumulative GHG emissions in every 365 day period since 1990 with a latency of less than two months. In other words, we produce daily estimates of annual emissions (where annual is thought of as a rolling 365 day window). Since we provide daily estimates of annual emissions, the estimate occurring on the last day of a given year is an estimate for the cumulative emissions that year. This figure is comparable to the annual value coming from the NZGGI for that year.

We compare a number of potential models in terms of prediction accuracy, model stability and over-fitting. The approach can be applied to any sector and has so far been applied to the Energy and Agriculture sectors of the NZGGI [34]. These sectors were chosen to begin with for their relative contribution to the NZGGI, making up 90.0% of emissions in 2020 [35].

The input variables for each sector are called *indicators*. There is a growing precedent that a relatively small number of indicators can be sufficient to model broad greenhouse gas emissions systems, recent examples including [12, 48]. Using this approach, we are able to model the Energy and Agriculture sectors with just four and six indicators, respectively. Our models do not assume a relationship between indicators and the emissions of their sectors. Thus, to explain the inner workings of the model we have to infer relationships between indicators and predictions by the model. We use subject knowledge to verify that these relationships are sensible. In addition, we ensure our methodology is repeatable between editions of the NZGGI.

We assess uncertainties carefully in terms of accuracy, stability and over-fitting in Sections 3.1.1 to 3.1.3. Additionally, we compare our uncertainties to typical inter-annual variation in Section 3.2 to mitigate the concern that the error due to our model exceeds expected changes from one year to the next. Results are derived in Section 3 using both of our high frequency and low latency of emissions reporting. The insights resulting from our models could improve current understanding of emissions in a landscape changing day to day.

2. Methods

2.1. **Overview.** We use cross validation to compare a number of potential models that predict NZGGI sector emissions using data on variables related to sector emissions. A model selection protocol based on prediction accuracy, model stability and over-fitting is used to decide a preferred model. We apply the preferred model to estimate the Energy and Agriculture sectors of the NZGGI [34]. These sectors were chosen to begin with for their size, together making up 90.0% of emissions in 2020 [35]. We call the independent variables used for each sector *indicators*, which we describe in detail in Section 2.2. Our models are non-parametric in the sense that we do not assume a data model. Thus we face the risk of selecting a model that cannot be explained. We combat this risk in Section 3.3.

The main objective of the model is to attain high frequency and low latency outputs. Because of limited data availability, to achieve the objective we make some simplifying choices: (1) as described above, our dependent variables are sector emissions (rather than estimating every source of emissions and then aggregating) and (2) we estimate carbon dioxide equivalents directly (rather than estimating each GHG type individually and then aggregating such as in [24, 48]).

The NZGGI updates annually. The latest edition at the time of writing reports until 2020, however our analysis is applicable to all editions of the NZGGI. Therefore, we discuss the model mostly with respect to any edition of the NZGGI. Any results presented pertain to the 2020 edition [2].

Code	Name	Unit	Frequency	Sector	Direct/indirect	Data sources
LEE	Liquid and electric energy	[1]	Daily	Energy	Direct	[1, 6]
REM	Emissions intensity from new vehicle registrations	[kgCO$_2$-e/km]	Monthly	Energy	Direct	[32, 5]
ETV	Emissions associated to national traffic volume	[kgCO$_2$-e]	Daily	Energy	Direct	[36, 32, 5, 6]
COA	Weight of coal production	[tonnes]	Quarterly	Energy	Indirect	[3]
COW	Number of dairy cattle	[number]	Annual	Agriculture	Indirect	[8]
MEA	Weight of meat produced at export-inspected facilities	[tonnes]	Weekly	Agriculture	Indirect	[4, 11]
FIM	Quantity of imported fertilisers	[tonnes]	Monthly	Agriculture	Indirect	[10]
EXM	Food exports for meat	[$]	Monthly	Agriculture	Indirect	[9]
EXD	Food exports for aquaculture and dairy	[$]	Monthly	Agriculture	Indirect	[9]
EXF	Food exports for horticulture	[$]	Monthly	Agriculture	Indirect	[9]

TABLE 1. Indicators used to model sectors. Direct means the indicator considers the emissions intensity of the underlying activity. Definitions of the indicators are given in Appendix A.

Existing research involves data sets of a wide variety of frequencies and time-spans. The most similar studies to ours in these terms are [27, 48, 51], which have an annual frequency and time-spans of 1990-2018, 1990-2019 and 1997-2017, respectively. We have a time-span of 1990-2020 when using the latest edition of the NZGGI. We collect new indicator data daily so that predictions using the model can be made daily. This means our outputs not only estimate year-end values that mimic the NZGGI but also interpolate between those year-end estimates and leave us with a low latency of less than two months in practice. A more detailed description of our daily estimates of emissions is given in Section 1.

2.2. **Indicators.** *Indicators* are times series associated with natural variables that are chosen as input variables for the model of a given sector. There are many variables that could serve as indicators of emissions in our model. The indicators discussed in this paper are non-exhaustive and this model could in fact be replicated with an entirely different set. We therefore do not claim they are the most important of the potential indicators, but they are sufficient in the sense that they effectively model their sectors at this point in time and are available dynamically. Our indicators are chosen for: (1) their systematic relevance to GHGs, (2) a resulting relationship with sector emissions predictions that is sensible, and (3) the availability of the necessary data. The indicators used in the model are summarised in Table 1.

Each indicator represents an emissions intensive activity (e.g. kgCO$_2$-e per km of petrol-fueled vehicle travel). Intensity (i.e. emissions per unit of activity) helps to account for the necessary dissociation of emissions with activity and productivity. If an indicator considers intensity, then we say it is *direct*. For example, one of our indicators measures the emissions potentials of new vehicles on New Zealand roads based on motive power, year of manufacture and gross vehicle mass. The data comes from a national vehicle register that is updated monthly. Therefore, our model accounts for variation in emissions intensity over time via indicators such as these. For more information, see Appendix A.2. If an indicator is not direct, we say it is *indirect*. We use a combination of direct and indirect indicators.

We check for updates to our indicator data on a daily basis to ensure the data is as up to date as possible. Where a dataset is not updated on a daily basis, the dataset is held constant until the present day and interpolated linearly between historic data points in order to provide a daily value. In the extreme case where a dataset is updated only annually, this would mean the data is held constant for up to 365 days.

In Appendix A, we describe the construction of each indicator in detail. Given the hundreds of sources of emissions featured in the NZGGI, the indicators presented in Table 1 appear scarce. We acknowledge there are many potential indicators, and as new datasets emerge, these may model the sectors better than those datasets currently being used. While the indicators used here are not

necessarily optimal, they are sufficient for this proof of concept. It is important to note, as previously discussed, that the benefit of this approach is the limited number of variables required within the model.
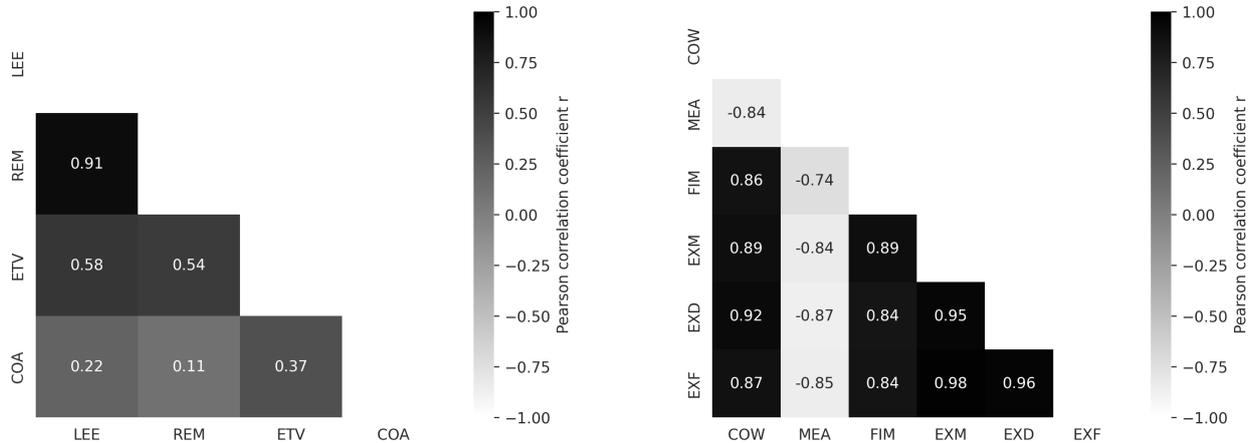
(A) Energy. LEE: Liquid and electric energy, REM: Emissions intensity from new vehicle registrations, ETV: Emissions associated to national traffic volume, COA: Weight of coal production.

(B) Agriculture. COW: Number of dairy cattle, MEA: Weight of meat produced at export-inspected facilities, FIM: Quantity of imported fertilisers, EXM: Food exports for meat, EXD: Food exports for aquaculture and dairy, EXF: Food exports for horticulture.

FIGURE 1. Heatmap of correlation matrix showing high collinearity between indicators in each sector.

Furthermore, the performance of regression models is generally sensitive to collinearity between indicators [18]. The scarcity of useful data makes it difficult to find indicators with low collinearity. We note that there exists high collinearity between indicators as shown in Figure 1. While methods exist to aggregate correlated variables, we keep our indicators as individual inputs for simplicity. Therefore, we will prefer machine learning models that have no such requirements around collinearity of indicators. Certain machine learning models also allow non-linear relationship between indicators and sectors [48]. This is valuable because including non-linear indicators can be critical for prediction accuracy [51].

We note that it is not intuitive that there is a negative correlation between MEA and EXM. However, there are possible explanations for this relationship, such as the export-inspected facilities where MEA is counted having a limited scope or the variation in the value of exports over time outweighing an actual decrease in quantity of meat exports. This requires further investigation. We discuss this further in Section 3.3.

2.3. **Model selection.**

2.3.1. *Models.* We compare three models in our selection process: bounded-variable least squares [45] (LS), random forests [14] (RF) and extra trees [21] (ET).

Bounded-variable least squares [45] (LS) solves the usual linear least squares problem with the added constraint that fitted coefficients lie within user-specified bounds. We use the Python programming language [49] for our analysis. In particular, LS is implemented with a linear least squares solver from the SciPy project [50]. For each indicator, we set the bounds to either $[0, \infty)$ or $(-\infty, 0]$ depending on the known systematic relationship between indicator and emissions. This aligns with the use of a 'relation expected sign' as in [23, Table 3]. [12] use such bounds to classify indicators as 'enablers' or 'de-enablers'. Significant statistical analysis is required to verify the use of LS, as demonstrated in [23]. As mentioned in Section 2.2, collinearity and non-linearity of indicators makes this infeasible in our setting. In selecting our optimal model, we will select against LS anyway because it is outperformed by the other models in terms of prediction accuracy (see Table 2). This suggests that the latent relationships are complex and/or non-linear. We therefore omit statistical details relevant only to LS.

Random forest [14] (RF) is an ensemble method that builds a collection of decision trees by incorporating a random component into their construction and makes a prediction by aggregating the predictions of each tree. The technique is *non-parametric* in the sense that RF does not assume a data model. However, there are various hyper-parameters for the structure of the forest and its trees.

These are manipulated in order to *tune* the model to optimise its predictions. This is discussed in detail in Section 2.3.2. We implement RF using the random forest regressor by Scikit-Learn [41].

We elaborate on the random component mentioned above in the construction of each tree to aid the description of the next model. Each decision tree in the forest involves a random sample of indicators and data points. In RF, these are sampled with replacement. In our application we have only 31 data points to begin with (each year from 1990 to 2020). Hence, we expect we cannot afford sampling with replacement in the random component described above. This motivates us to use ET.

Extra trees [21] (ET) is similar to RF but is different as follows. While the trees in RF involve sampling with replacement, the trees in ET involve sampling without replacement. Thus each tree in ET is likely to learn from a more comprehensive sample of data. In a similar study to ours [27], RF did not perform exceptionally well, placing fourth among eight models considered. We show in Section 3.1.1 that our implementation of RF is more accurate than [27]. We observe in Section 3.1.2 that stability issues of RF are resolved by our inclusion of ET. We use the extra trees regressor by [41] to implement ET.

2.3.2. *Cross validation.* *Over-fitting* occurs when a fitted model performs worse on unseen data than on its training data. Cross validation is a technique used to assess for over-fitting. We follow a standard workflow that incorporates hyper-parameter tuning when applicable.

Since the tree-based methods have hyper-parameters to be tuned, cross validation occurs at two levels: (1) validating to find the best hyper-parameters and (2) testing to ensure the model is not over-fitting after retraining on the best hyper-parameters. Thus we use a workflow that is best described as 'repeated double cross validation' by [19] (see also [25, Figure 7.4]). This means that at the lowest level the data is split into three parts: pre-training, validating and testing. We use the train-test split utility by [41] with a test size of 6 out of the 31 data points. This gives a train:test ratio of about 80:20, which is the same ratio as in [51]. If the model on hand is RF or ET, then hyper-parameters need tuning. Randomised search through a given hyper-parameter space is used to find the best hyper-parameters. We use randomised search instead of its predecessor, grid search, because randomised search is known to be better at navigating situations where hyper-parameters have differing importance [13, Figure 1]. Randomised search cross validation by [41] with 5 folds implements the randomised search, which repeatedly splits the 25 training data points into 20 for pre-training and 5 for validating.

The hyper-parameter space used both for RF and ET is defined with the ranges:

- Number of trees: $\{200, 201, ..., 219\}$,
- Maximum depth of tree: $\{1, 2, 3\}$,
- Minimum number of samples required to split: $\{2, 3, 4\}$, and
- Minimum number of samples required to be at a leaf node: $\{2, 3, 4\}$.

These were chosen by running a large number of investigations to find a large hyper-parameter space in which the models were stable in the sense we describe in Section 2.3.3.

Fitting each model on a particular train-test split of the data (as described above) defines a single simulation in our cross validation. We run 30 such simulations. This makes our cross validation more comprehensive than [27, 48], which consider only one simulation each.

2.3.3. *Model outcomes.* The objective is to select the optimal model in terms of certain 'model outcomes'. The model outcomes used in [16] are: over-fitting, prediction accuracy and execution time. The bounds on the parameters, described earlier, ensure that the use of the attributes remains aligned with any latent data structures. Execution time is not important to us since our data sets are small, so we omit this outcome.

A novelty in our approach is to include an outcome about the stability of the structure of the model. Different models have different structures, so the notion of stability has to adapt to each model. For LS, we consider the distribution during cross validation of the fitted coefficients in the linear model. For RF and ET, instead of coefficients, we consider the values of indicator importance in the tuned estimator. A model is considered more stable if it has a smaller variation in the distribution above. These are described in more detail in Section 3.1.2 and pictured in Figure 2.

Thus, our objective is to find the optimal model in terms of the model outcomes:

(1) Prediction accuracy, in terms of mean absolute percentage error (MAPE):

$$\frac{100}{n} \times \sum_{x \in X_{\text{test}}} \left| \frac{x - \hat{x}}{x} \right|$$

where $\hat{x}$ is the prediction for each x in a given test set X_{test} of size n in cross validation, like in [16],

(2) Model stability, in terms of distribution of fitted coefficients/indicator importance as described above, and

(3) Over-fitting, in terms of the average difference of MAPE between training and testing data, like [16].

3. Results

3.1. **Framework to assess models to predict sector-level greenhouse gas emissions in national inventories.** Cross validation was applied to compare a range of models of greenhouse gas emissions by sector in terms of a comprehensive list of model outcomes (Section 2.3): prediction accuracy, stability and over-fitting. A machine learning model, extra trees (Section 2.3.1), was found to perform the best according to our selection protocol (Section 3.1.4). Relationships between sector emissions and indicators were inferred from the model (Section 3.3), and subject knowledge was used to explain these relationships. Outliers in the relationships were identified via Mahalanobis distance and explained in context.

3.1.1. *Prediction accuracy.* Statistics for prediction accuracy are given in Table 2, which shows ET is optimal across both sectors.

sectors model statistics	Agriculture			Energy		
	LS	RF	ET	LS	RF	ET
count	30	30	30	30	30	30
mean	1.962	1.484	**1.347**	3.818	3.565	**3.211**
std	0.560	0.634	0.622	0.837	0.911	0.866
min	**0.444**	0.448	0.526	2.451	1.732	**1.602**
25%	1.594	1.162	**1.071**	3.356	3.167	**2.732**
50%	1.999	1.406	**1.314**	3.664	3.617	**3.194**
75%	2.244	1.635	**1.455**	4.452	4.023	**3.779**
max	**3.199**	3.844	3.967	5.612	6.232	**5.583**

TABLE 2. Descriptive statistics for MAPE (%) values in test set for each of the 30 simulations. LS: bounded-variable least squares, RF: random forest, ET: extra trees. A value for a summary statistic (e.g. mean) is in bold if it is optimal compared to other models in the same sector. This gives an indication of which model is optimal while accounting for the distribution of MAPE values observed throughout simulation.

Our results are comparable to those found in [27] due to the similarity in methods and initial data. In both settings, MAPE is used to measure error when predicting annual values of CO_2-e. Our models have resulted in higher accuracy than those presented [27]. The comparable model to LS in [27] is 'ElasticNet' since it too is a regularised linear model. The comparable model to both RF and ET in [27] is 'Random Forest'. LS achieves a median MAPE of 2.0% in Agriculture and 3.7% in Energy, whereas ElasticNet achieves a MAPE of 19.462% in one application of the model and 41.619% in another [27, Tables II and III]. ET (respectively, RF) achieves a median MAPE of 1.3% (respectively, 1.4%) in Agriculture and 3.2% (respectively, 3.6%) in Energy, whereas Random Forest in [27] achieves a MAPE of 9.470% in one application and 21.522% in another [27, Tables II and III]. Our higher accuracy is possibly due to our use of hyper-parameter tuning. There is no discussion of tuning in [27].

3.1.2. *Model Stability.* Recall from Section 2.3.3 that LS (respectively, ET and RF) is considered to be more stable if it has smaller variation in the distributions of its fitted coefficients (respectively, indicator importance values). These distributions are pictured in Figure 2. We observe stability can vary in a number of ways.
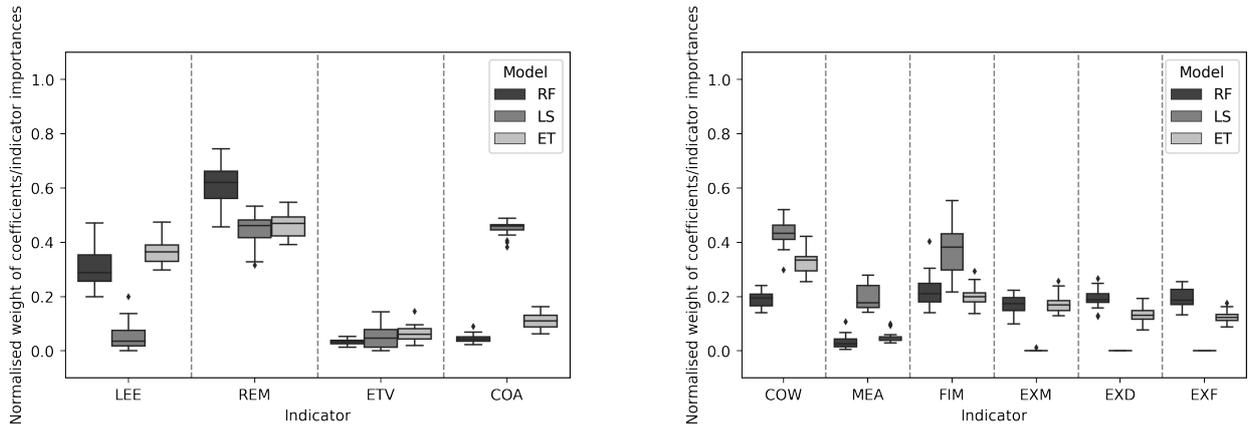
(A) Energy: LEE: Liquid and electric energy, REM: Emissions intensity from new vehicle registrations, ETV: Emissions associated to national traffic volume, COA: Weight of coal production.

(B) Agriculture: COW: Number of dairy cattle, MEA: Weight of meat produced at export-inspected facilities, FIM: Quantity of imported fertilisers, EXM: Food exports for meat, EXD: Food exports for aquaculture and dairy, EXF: Food exports for horticulture.

FIGURE 2. Box plots picturing distribution of fitted coefficients (for LS, bounded-variable least squares) and indicator importance (for RF and ET, random forest and extra trees, respectively) values as a measure of stability of each model.

Firstly, the order of indicators by median value depends on the model. For example, in RF, the median value for FIM (Quantity of imported fertilisers) is slightly higher than that of COW (Number of dairy cattle), whereas in ET the box and whiskers for COW is entirely above the upper quartile for FIM.

Secondly, notice simply how long the boxes and whiskers in some models are, and therefore how unstable the models can be. ET has consistently short boxes and whiskers. On the other hand, in RF for Energy, the length of the box and whiskers for the two most important indicators are long, especially in comparison to the box and whiskers for the same indicators under ET. This instability in RF is undesirable since it demonstrates high uncertainty about the impact of each indicator in the model.

Lastly, models can be more or less 'polarising' about the coefficient/importance of an indicator. Notice LS is very polarising since it tends to assign either very high or low coefficients. This is particularly evident in the Agriculture sector (see Figure 2b); three of the six indicators' coefficients are always 0. This annihilation of indicators is a common feature of a LS because of the bounding of variables. This is likely to be a manifestation of collinearity, highlighting the importance of variable stability as one of our model outcomes.

In summary, we observe in Figure 2 that ET is the most stable compared to RF and LS.

3.1.3. *Over-fitting.* LS had the least over-fitting compared to ET and RF. That is, the maximum difference in MAPE between training and testing in LS was 1.9% compared to 3.1% for both ET and RF in Agriculture and was 2.8% compared to 4.2% for RF and 3.6% for ET in Energy (see Table 3). That said, all models performed well as demonstrated by means in Table 2. These values are acceptable in that similar or higher values of over-fitting are expressed in [16, Table 4].

Recall from Section 2.3.2 that there are only 31 objects in our dataset. It is well known that over-fitting is more common for smaller samples. Yet we successfully mitigated over-fitting as above. Moreover, aside from [16], we found it is not common in the literature to explicitly analyse over-fitting at all. For example, [37, 28, 27, 48] do not mention over-fitting. The study [51] mentions over-fitting and the use of training and testing sets but does not report any measures of over-fitting.

3.1.4. *Selection.* Our selection protocol is as follows: given the models LS, RF and ET, remove the model performing worst in terms of prediction accuracy; from the remaining two models remove the model performing worst in terms of model stability; the remaining model is selected provided it is not over-fitting.

Over-fitting is regarded as a lower priority compared to accuracy because prediction accuracy is measured on the test set, so it mitigates over-fitting implicitly. Also, over-fitting is regarded as a lower

| sectors | Agriculture | | | Energy | | |
model statistics	LS	RF	ET	LS	RF	ET
count	30	30	30	30	30	30
mean	0.517	0.696	0.561	0.804	1.743	1.264
std	0.677	0.662	0.620	1.006	0.994	0.999
min	-1.339	-0.336	-0.324	-0.835	-0.232	-0.594
25%	0.093	0.344	0.255	0.239	1.379	0.748
50%	0.584	0.603	0.537	0.617	1.873	1.223
75%	0.938	0.894	0.714	1.605	2.318	1.881
max	1.865	3.126	3.069	2.834	4.157	3.639

TABLE 3. Descriptive statistics for MAPE difference (%) between test set and train set for each simulation. LS: bounded-variable least squares, RF: random forest, ET: extra trees.

priority than stability since an unstable model cannot be explained. In Section 4.1.5, we compare our selection protocol to that in [16].

We apply the selection protocol now. The measures in Table 2 demonstrate LS is the least accurate, so we remove LS from the running. The accumulated length of RF box and whiskers (Figure 2) is greater than that of ET, so RF is removed for being less stable than ET. The remaining model is ET and we recall that its over-fitting values were acceptable.

3.2. **Uncertainty in model compared to inter-annual variation.** Throughout Section 3.1 we recorded uncertainties of our models from the point of view of the model outcomes established in Section 2.3.3. In particular, this involved optimising prediction accuracy by minimising residuals. However, no matter how small the residuals are, an outstanding concern is whether the residuals exceed typical levels of inter-annual variation of NZGGI emissions estimates, in which case the predictions would be useless since the accuracy of the predictions would be outweighed by expected variation between years.

We have mitigated this concern by comparing our residuals to inter-annual variation (measured in terms of year-on-year differences in NZGGI emissions estimates). For both sectors, the maximum, upper quartile, median and lower quartile values are all significantly lower for our residuals than for inter-annual variation. The following values are given as percentages of the average value for the sector. For the Energy sector, the lower quartile, median and upper quartile for our residuals are 0.54%, 1.73% and 2.89%, respectively, whereas they are 1.20%, 4.08% and 5.04% for inter-annual variation. For the Agriculture sector, the lower quartile, median and upper quartile for our residuals are 0.26%, 0.54% and 1.27%, respectively whereas they are 0.47%, 1.00% and 1.71% for inter-annual variation. That is, we find the uncertainty of the model tends to be significantly below typical inter-annual variation.

We also consider the following year-wise paired view comparing inter-annual variation to our residuals. For each year, we divide the relevant value of inter-annual variation with our residual that year, and compute the median of these fractions. The median for the Energy sector was 1.804 (meaning inter-annual variation tends to be 80% higher than our residuals), while the median for the Agriculture sector was 1.322 (meaning inter-annual variation tends to be 32% higher than our residuals).

3.3. **Explanation of model.** Our selected model, ET, does not presume a data model. Hence, in order to check the model is sensible/can be explained, we must infer relationships between inputs and outputs a posteriori. A relationship is *inferred* between an indicator and its sector if a multivariate trend is present. We find every indicator has an inferred relationship with its sector. Each inferred relationship is tested for outliers by calculating Mahalanobis distances [22] and calculating associated p-values on a χ^2-distribution with 1 degree of freedom. Then, inferred relationships and their outliers are explained using subject knowledge.

Inferred relationships must be understood in the context of the indicators used. Changing the selection of indicators would lead to variation in the inferred relationships. In this way, our methodology has a limited capacity to make precise/absolute claims about drivers of emissions. With this in mind, we give two examples of inferred relationships and their explanations.

The indicator EXF (Food exports for horticulture) for Agriculture has a logarithmic trend (see Figure 3a). This is to be expected in horticulture since smaller farms are known to have higher average costs, which are associated to emissions. Thus, as production increases, we would expect emissions to increase at a high rate initially and then at a lower rate for higher levels of production.
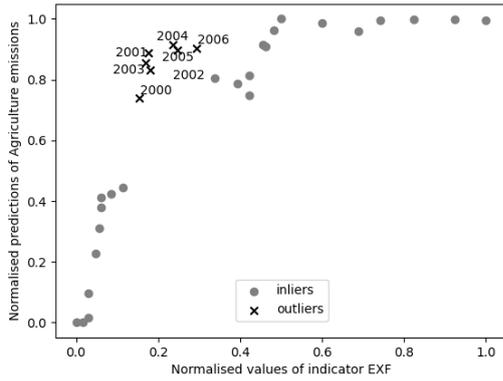
(A) EXF: Food exports for horticulture
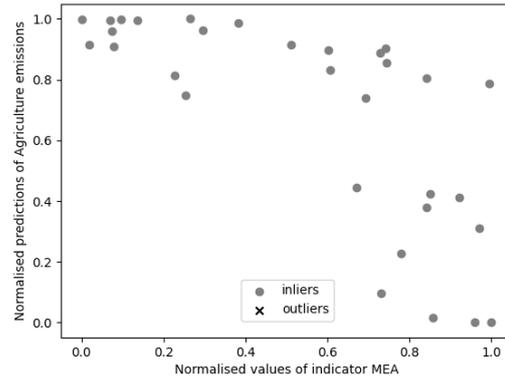
(B) MEA: Weight of meat produced at export-inspected facilities

FIGURE 3. Inferred relationships between indicators and predicted emissions. Outliers are identified and labelled with their year.

Outliers in this inferred relationship occurred between years 2001 and 2006. This coincides with a surge of Agriculture emissions. This surge is largely due to enteric fermentation in non-dairy cattle (category code 3.A.1 in the NZGGI [34]). However, EXF is not related to non-dairy cattle, so it is natural that the indicator could have as low values as it does in spite of the surge.

All indicators had unsurprisingly positive inferred relationships with emissions except for the following example. On average, predicted Agriculture emissions decreased while MEA (Weight of meat produced at export-inspected facilities) increased (see Figure 3b). Since every animal is a source of emissions, if animals are removed from the stock at a higher and higher rate (in order to support higher levels of meat production), this must lead to a reduction in emissions with all else being equal. Indeed other factors will come into play in order to support the continuous turnover of the stock. Recognising such other factors would be the role of other indicators like COW, which counts the total number of dairy cows.

Therefore, we stand by the negative relationship observed between MEA and emissions in the Agriculture sector. We also emphasise that we do not imply there is a direct causal relationship between animals' lives being ended and a decrease in emissions. Rather there is an explainable relationship that should be viewed not in isolation but in the context of other interactive factors in the model. In this way, the negative relationship between Agriculture emissions and meat production can be explained. No outliers were identified in this inferred relationship.

3.4. **Preliminary forecast of New Zealand Greenhouse Gas Inventory.** We observe similar trends between our emissions estimates and the NZGGI (see Figure 4). However, our model underestimates the magnitude of peaks and troughs in the Agriculture sector and to a lesser degree in the Energy sector. This is a known weakness of machine learning in predicting extremes [28].

We estimate Agriculture emissions in 2021 were $39,286$ ktCO$_2$-e compared to $39,299$ ktCO$_2$-e in 2020, which is a decrease of about 0.03%. We estimate Energy emissions in 2021 were $32,388$ ktCO$_2$-e compared to $32,069$ ktCO$_2$-e in 2020, which is an increase of about 0.99%.

3.5. **Energy emissions in 2020 COVID-19 lockdown.** We estimate total Energy emissions were $32,576$ ktCO$_2$-e in 2019, higher than the 2020 figure above, but only by about 1.58%. This comes in spite of several COVID-19 lockdowns. Indicators in our model for Energy appeared to recover from their shocks too quickly for lockdowns to have lasting impact for the year of 2020.

For example, consider transport (which accounted for 41.88% of Energy emissions in 2020 [35]); changes in state highway traffic volumes at selected telemetry sites across the country–a component of indicators in our model for Energy–were closely related to changes in COVID-19 alert levels (see Figure 5). Thus, traffic volumes recovered before the end of the year, as did emissions in the Energy sector.
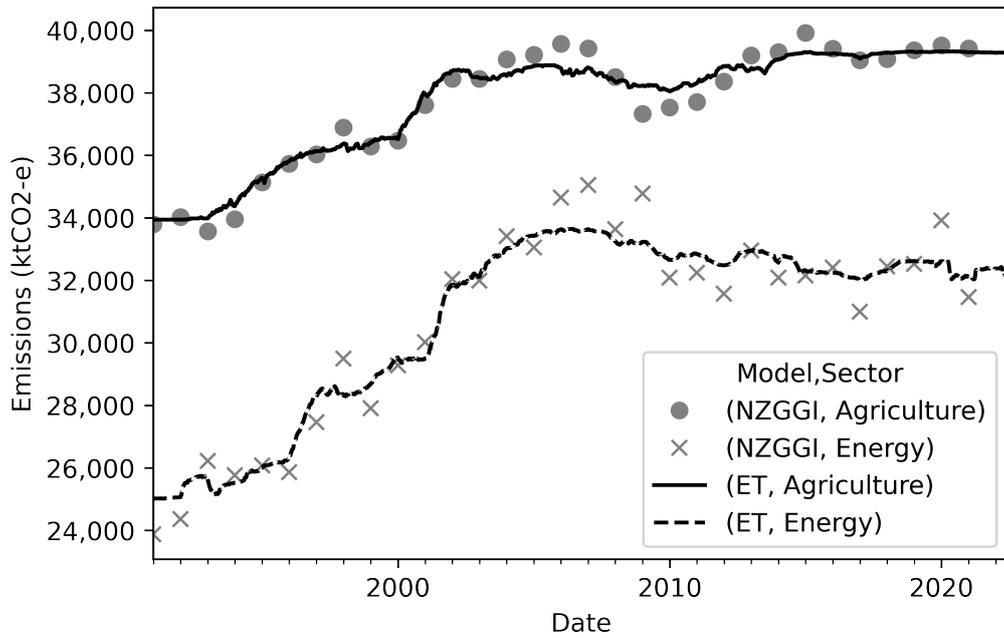
FIGURE 4. Emissions estimates for each sector modelled in our analysis according to NZGGI (the actual data) and extra trees–ET (the predicted data). This demonstrates the greater timeliness of our estimates both by being more frequent and having a lower latency (nowcasting).
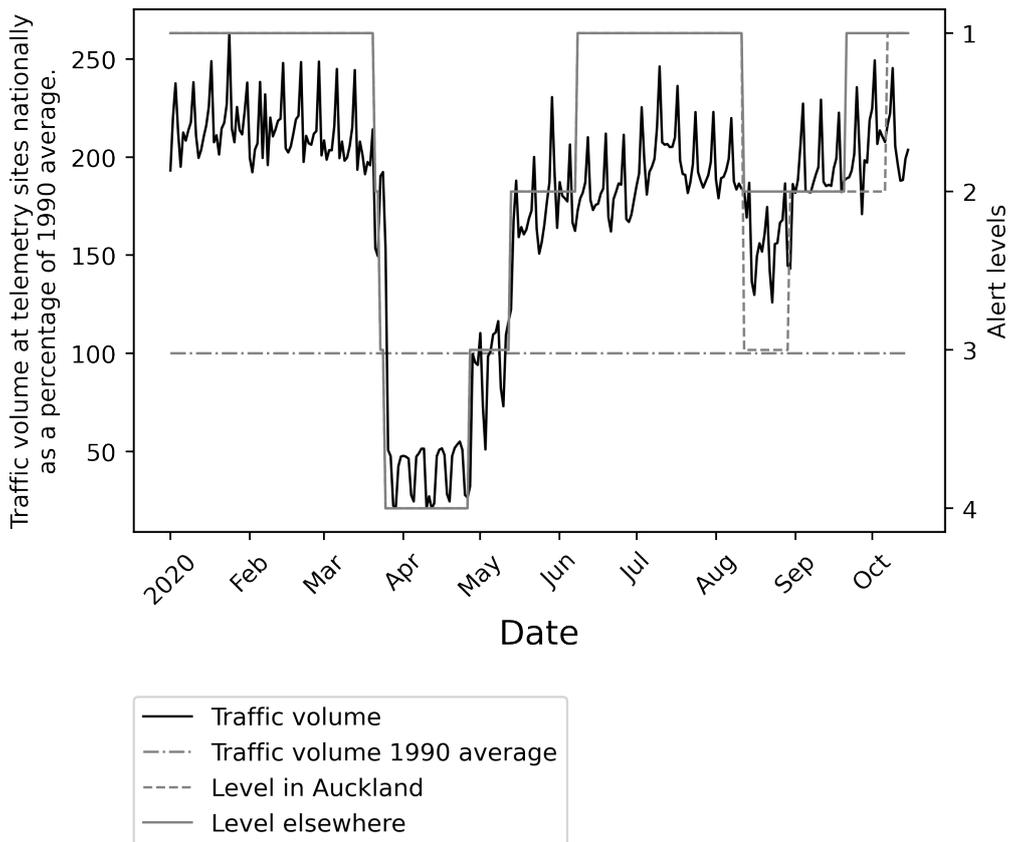

FIGURE 5. Overlay of COVID-19 alert levels in Auckland and elsewhere in New Zealand with daily national traffic volume at selected sites across the country. The traffic volume is expressed as a percentage of the 1990 average. It is involved in the following indicators: LEE (Liquid and electric energy) and ETV (Emissions associated to national traffic volume).

3.6. **Link between emissions and economic productivity.** Because we maintain our indicator data on a daily frequency, our model can make predictions on a daily basis of emissions occurring during 365 day windows, which aligns with the NZGGI on 31 December each year. An advantage this gives us is visualising more granular trends than if we only had annual predictions. We demonstrate the value of this through an example.

An historical comparison to the Global Financial Crisis (GFC) suggests a continued link between emissions and economic productivity. Observe in Figure 6 that the steep descent of emissions during the GFC (beginning in 2008) has a similar gradient to that of COVID-19 lockdown from 25 March to 26 April 2020. Given both the GFC and COVID-19 lockdowns are associated to significant economic impact, this raises the question of how exactly emissions is associated with economic activity.
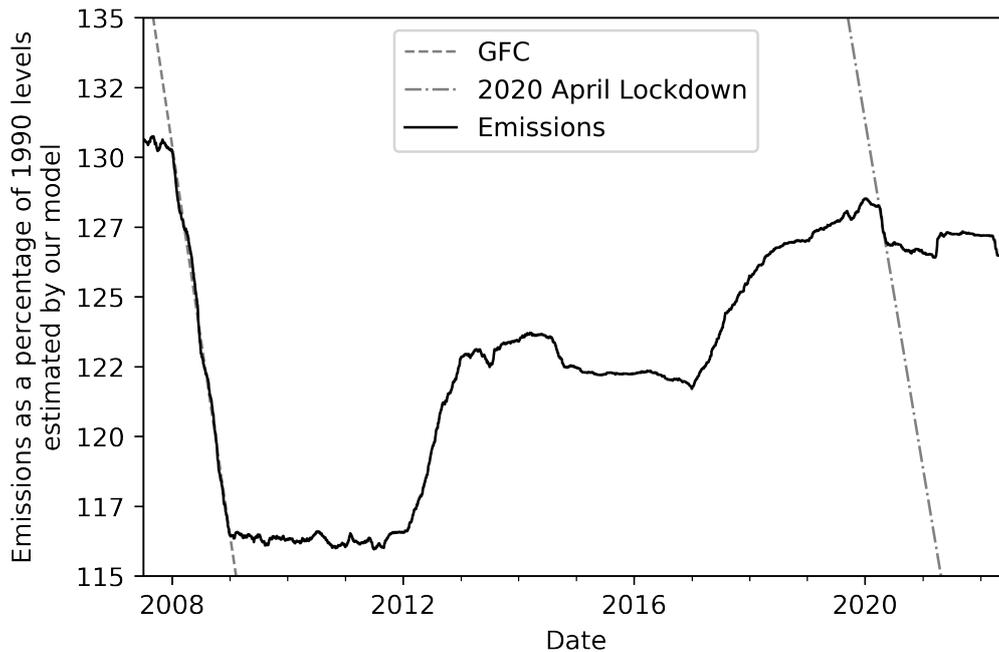

FIGURE 6. Net emissions as a percentage of 1990 levels with lines of best fit during the global financial crisis (GFC) and the COVID-19 lockdown from 25 March to 26 April 2020. This demonstrates the slope during the GFC (decrease of 1% every 26.2 days) is similar to that of the 2020 April lockdown (decrease of 1% every 29.5 days). In this sense, the GFC and 2020 April lockdown had comparable effects on net emissions.

The release of emissions budgets by the New Zealand Government in 2021 highlighted a turning point for the dissociation of economic activity and emissions. Budgets can be monitored closely with our dynamic estimates.

4. DISCUSSION

4.1. **Strengths.**

4.1.1. *Meeting the needs of nowcasting.* The outputs of the model are daily predictions of cumulative emissions in 365 day periods per sector, which coincide with the national inventory at the end of each year. This comes with a time lag of about two months given our current indicator data. Therefore, our model addresses four of five needed directions of improvements to emissions statistics demanded in [26]: time lag, frequency, sector-level activity and accuracy (discussed in Section 3.1.1 among our other model outcomes). The fifth direction is coverage of sources of emissions, which we discuss in Section 4.2.4. Thus, we meet a growing demand for nowcasting (see references in Section 1).

4.1.2. *Repeatability.* Our emissions estimates anchor to and complement the NZGGI. A key strength is repeatability for every new edition of the NZGGI. We hope this continuity of information on emissions will hold policies and actions to account.

4.1.3. *Comparison to methodologies using bespoke inventories.* Innovative techniques have been developed to investigate COVID-19 impacts on daily emissions [20, 30, 31, 39, 43, 54] with impressive results (e.g. daily emissions with a regional component separating CO_2 and air pollutants [24]). However, these studies are limited by depending on bespoke and labour-intensive daily inventories of emissions. To mitigate the labour, they compensate in various ways.

Firstly, the studies focus on short windows of time, whereas our work reports from 1990 up to two months prior to the present day. Greenhouse gas emissions systems change gradually, so our long-term perspective is valuable.

Secondly, scaling/repeating methods in these studies would be costly since it would require curating another bespoke inventory. In particular, scaling the methods to a national inventory like the NZGGI is hard to foresee; the 15 month delay before each edition of the NZGGI is released suggests the annual NZGGI is labour-intensive enough.

Lastly, the daily inventories used in these studies try to reproduce third-party estimates of countries' national emissions, rather than official estimates. This leads to a fragmented picture of emissions. Our methodology aligns with the existing national inventory.

4.1.4. *Local quarterly emissions reports.* Statistics New Zealand (StatsNZ) has begun experimental reporting on quarterly emissions [47]. We comment on distinctions between our methodologies here.

StatsNZ reports are more dynamic in one way than ours, estimating emissions occurring within any given quarter, whereas our estimates are always for 365 day periods. However, our estimates are more dynamic than theirs in the sense that we provide daily predictions (of annual periods). Lastly, StatsNZ reports currently involve a different classification of 'industries' compared to the 'sectors' in the NZGGI.

4.1.5. *Novelty of our approach.* The review by [17] of 776 articles on emissions predictions suggests tree-based models are not the most widespread. We did find a handful of studies using tree-based models for emissions predictions [16, 27, 28, 29, 51, 52], but their applications vary from ours. Hence, our research is novel in terms of the application of tree-based models to estimating the NZGGI. Furthermore, the performance of our methods suggests that techniques that can robustly handle data that potentially violates the assumptions of regression are most suitable, such as random forest and extra trees.

The most similar studies to ours by methodology and application were [16, 27, 48]. We remark on points of difference that are strengths of our approach.

Re-training our model day after day is operationally necessary to incorporate the latest data and give the latest estimates. This raises a question of the stability of the structure of the model between training, which was assessed in Section 3.1.2. The comparison studies do not assess stability.

Our selection protocol from Section 3.1.4 is decision-driven instead of quantitatively-driven as in [16, Figure 5], which is more definitive but less flexible than ours. That is, [16] only accounts for average performance, whereas our protocol allows us to take into account the distribution of performance throughout simulations. This is important because model selection has its subtleties; regarding overfitting, [38] provides evidence that finding the optimal model is not the same as finding the model with the least over-fitting. We discuss a blend of the protocols in Section 4.2.1.

Recall our selection protocol is based on three model outcomes: accuracy, stability and over-fitting. Cross validation is used to gather information these outcomes. Compared to similar studies [27, 48], our analysis and results:

(1) Employ a more thorough cross validation approach (Section 2.3.2),
(2) Demonstrate higher accuracy (relative to contemporary methods referenced in Section 3.1.1), and
(3) Assess the extent of over-fitting (Section 3.1.3).

4.2. **Limitations.** Limitations are considered in three types:

- Methodological limitations: ways we are limited due to our approach in theory,
- Operational limitations: limited availability of data, and
- Explanatory and predictive limitations: limitations on the how well the model explains the system at hand and on the predictions we can make with the model.

4.2.1. *Methodological limitations.* A key limitation of the methodology is that it treats high level aggregates directly by taking sectors' emissions as dependent variables. We could enhance the model by partitioning sectors whenever we find that a source can be modelled separately into 'the source' and 'not the source' parts. The extreme case of this approach would be to model every source and sink individually.

Also, the methodology estimates CO_2-e variables directly. Modelling each greenhouse gas type is of interest from both a policy and practical perspective.

Our model estimates the NZGGI, which carries the imperfections of the IPCC guidelines it follows [53]. Thus, our model limited in every way the NZGGI is limited.

As described in Section 4.1.5, the protocol we use to select our preferred model is less definitive than in [16]. In order to make our protocol more definitive but maintain a view of the variation of model performance between simulations, future research could investigate the following approach: define an overall performance score by computing differences between upper and lower quartiles for each model outcome (accuracy, stability and over-fitting), multiplying the difference by the mean of the model outcome when the model outcome is supposed to be minimised (accuracy and stability), and taking a weighted sum of these components. The optimal model for a given sector would be the model that minimises this score for that sector. If different sectors have different optimal models, the score could be broadened to aggregate over sectors in order to determine a general optimal model.

4.2.2. *Operational limitations.* Data availability is the main limitation to the operational quality of the model. In particular, limited availability leaves us with a latency of two months. For this reason we are always investigating new sources of dynamic and near-real-time data to improve the model.

4.2.3. *Explanatory and predictive limitations.* Whilst our model selection prioritises predictive power over explanatory power [14], we ensure that our models are explainable. To this end, we (1) selected indicators with known systematic relationships with emissions and (2) verified that the relationships between indicators and predictions were sensible. Future research will involve increasing the explanatory power of the model.

In order for prediction accuracy statistics in Table 2 to apply to future estimates, we assume there is no change from one NZGGI edition to the next. We measure how bad an assumption this is by the extent of the changes between editions. For example, from the 2018 edition to the 2019 edition in the Agriculture sector the emissions for each year reported on changed by between 4.2% and 5.2%. Whereas from the 2019 edition to the 2020 edition in the Energy sector, the reports changed by between -0.5% and 1.6%. This presents a moving target for our model. Estimating compound uncertainty due to these NZGGI amendments is non-trivial because amendments are non-uniform as noted above; reports for some years are amended to be higher while some years are amended to be lower. This dependence on the inventory at hand is a common disadvantage across any model estimating an inventory that updates over time. However, the advantage of using an updating inventory is the longevity of the applicability of our model. That is, since our model can be applied to any edition of the NZGGI, we continually revise our analysis according to new editions of the NZGGI each year.

4.2.4. *Indicator limitations.* The quality of any model depends on the quality of the data to which it is fit. Therefore, it is important to note limitations of our indicators. Their constructions are described in Appendix A. They have a limited coverage of the many sources of emissions, and coverage is an important direction for improvement of emissions information [26]. Screening a larger number of indicators when data becomes available will be beneficial for the coverage of our model, and a formal indicator selection such as in [51] will be necessary.

Many indicators do not account for emissions intensities. Multiplying indicator values by a coefficient from an IPCC look-up table is not useful because models tend to be mostly scale-independent. Thus, a time-dependent analysis of emissions intensity for the given activity is necessary.

We present a preliminary model, with a limited number of indicators covering only Energy and Agriculture. We recognise the importance of dynamically modelling the remaining sectors of the NZGGI. Fortunately, the majority of the work has been done, resulting in a generic sector model. Modelling the remaining sectors only requires trialling new data sources when they become available and assessing their outcomes.

5. Conclusions

This study develops a proof of concept that there is a framework by which machine learning can be used in combination with a small number of indicators to model sector-level emissions estimates at a higher frequency and lower latency to complement the New Zealand Greenhouse Gas Inventory (NZGGI) and any other relevant methodologies currently investigating greenhouse gas emissions in New Zealand. Estimates are reported every day for the cumulative emissions in the 365 day period ending that day, which gives a more continuous stream of emissions statistics with a time lag of approximately two months.

The high frequency allows us to compare shocks that take effect in time periods shorter than one year. For example, our results show the COVID-19 lockdown from 25 March to 26 April 2020 in New Zealand coincided with a rate of decrease in emissions similar to that of the global financial crisis of 2008.

The low latency allows us to forecast, up to 2021, the latest edition of the NZGGI (reporting only up to 2020) to estimate total emissions in the Agriculture and Energy sectors. We estimate Agriculture emissions in 2021 were $39,286$ ktCO$_2$-e and Energy emissions in 2021 were $32,388$ ktCO$_2$-e, which represents a decrease of 0.03% and an increase of 0.99%, respectively, since 2020.

In the face of rapid policy and technological change, these insights based on our model can improve current understanding of emissions in a landscape changing day to day.

We evaluate uncertainties in our methodology and performance of a number of models in terms of prediction accuracy, stability, over-fitting, and by comparison to typical inter-annual variation. This assessment is more rigorous than contemporary studies in many ways. The selected model carries an error margin of 1.314% for Agriculture and 3.194% for Energy (median during simulation study). The model is explainable in the sense that indicators used have systematic relationships with the emissions in their sectors, and the model fits sensible quantitative relationships between each indicator and the predictions it makes. That said, the model is not designed to tell us which indicators of emissions are stronger drivers than others.

This study has highlighted the demand for the release of relevant data sets in a more timely fashion. More granular data would lead to significant improvements to the model. We also emphasise the need for the continuation of a bottom-up emissions model as presented in national greenhouse gas inventories; this would provide a test-bed for a more dynamic approach and to directly measure the role of technological advances across the broad spectrum of emissions.

We note that although methodologies provided in this paper were able to estimate sub annual inventory style estimates within New Zealand, transferring this methodology to other countries would prove challenging, due to the nature of data used in this analysis.

Our methodology and emissions estimates will continue to be updated and developed to assess the ongoing variability in New Zealand's emissions. Further research will be undertaken for various purposes as discussed in Section 4, not least of which in order to estimate emissions by gas type, providing direct comparison against progress towards split gas targets, and regional emissions estimates that support more direct action at a local level. We also plan to expand our estimates to other sectors of the NZGGI, although note that our current methodology dynamically reports on sectors making up 90.0% of New Zealand's emissions.

Declaration of competing interest

Data availability

The data used in this research is either publicly available via the cited source or the data was made available to the researchers through a data commercialisation agreement with one of the data partners of DOT Loves Data. The data provided by the data commercialisation agreement is therefore unavailable to access.

Acknowledgements

literature and methodological support. We would also like to thank Louisa Howse for their suggestions in the early stages of method development, along with those at Transpower and emsTradepoint for their continued interest. The project has been jointly funded by Transpower, Callaghan Innovation and DOT Loves Data.

Appendix A. Indicators

Table 1 collates details of indicators used as independent variables in sector models. Descriptions of these indicators are given in the following subsections.

A.1. **Liquid and electric energy.** This indicator comprises of an aggregate of traffic volumes and the estimated carbon emissions due to national electricity generation [1]. Each component is described below. The increasing rate of electric vehicles draws a systematic connection between the components, which partly motivates their aggregation.

The components are normalised and added together with equal weight to give 'Liquid and electric energy'.[2] Equal weighting is chosen to give the model equal view of both variables since we are interested in trends rather than magnitude of sources of emissions.

Traffic volumes variable. Waka Kotahi New Zealand Transport Agency provides an application programming interface for daily traffic volumes at sites around the country [6] dating back to 2018. We restrict our attention to telemetry sites because of their accuracy and frequency. We extend the data back to 1990 by matching it with annual historic records in pdfs and xlsxs [7]. This gives a unified record of traffic volumes with many site descriptions. A difficulty is that the description of a given site can change over time. To connect the traffic volumes per site in spite of changing descriptions, we built an adjacency table between the site descriptions and a manual selection of keywords (e.g. 'Ngauranga', 'Drury' and 'Taupiri') where a keyword is adjacent to a site description if the key word appears in the site description. Sites being established or disestablished over time could lead to over- or under-estimates, respectively, of the increase in traffic volumes over time. For this reason we restricted to only those sites that have run continuously from 1990 to the present day. The traffic volume variable is the time series given by summing the series for each of these sites.

Electricity generation variable. The estimated carbon emissions due to national electricity generation is defined in the em6 application programming interface integration guide [1].

A.2. **Emissions intensity from new vehicle registrations.** This indicator captures the emissions potential of New Zealand's vehicles based on motive power, year of manufacture and gross vehicle mass. Previous studies have highlighted the importance of including a view of the country's vehicle fleet [23].

Vehicles in the Motor Vehicle Register [5] are assigned an emissions potential ($kgCO_2$-e/km) based on locally-approved IPCC-developed emissions factors [32]. The emissions factors consider motive power ('petrol', 'diesel', 'petrol hybrid', 'electric', 'lpg', 'plugin petrol hybrid', 'petrol electric hybrid') and vehicle year, assuming a gross vehicle mass of 2000-3000kg. We assign to each vehicle the relevant emissions factor after adjusting for its actual gross vehicle mass. The adjustment applies the principle that fuel economy is proportional to vehicle mass, which is comparable to the proportions-based approach to understanding environmental impact of vehicles in terms of their mass [42]. Other studies use a congestion index [31] and may match traffic flow data with a vehicle database to differentiate gasoline from diesel powered vehicles [24]. Our view has an advantage over these methodologies by differentiating vehicles by more motive power categories as well as by vehicle age and gross vehicle mass.

A.3. **Emissions associated to national traffic volume.** This indicator is the product of the emissions potential of the fleet and the traffic volumes variable from Appendix A.1. Estimating the emissions potential of the fleet involved estimating the half-life of a vehicle on New Zealand roads and using this in an exponentially weighted mean over newly registered vehicles each carrying their own emissions potential.

Recall our indicator of emissions intensity of vehicles registered onto New Zealand roads from Appendix A.2 based on the New Zealand Motor Vehicle Register (MVR). A limitation of this indicator

---

[2]This aggregation was undertaken initially to avoid collinear indicators when we were considering only models that demanded this. Since the introduction of the tree-based models, it is now feasible to remove this aggregate, which we will undertake in future research.

is that we do not know how long these vehicles spend on the road. There does exist a dashboard tool that provides a 'vehicle status' feature for the MVR at an aggregated level, which tells whether a vehicle is currently active or inactive. However, the API that provides the MVR to us vehicle-by-vehicle does not include 'vehicle status'. Therefore, this feature was not readily available to us in order to calculate how long vehicles spend on the road. Instead we opted to use more readily available data for this purpose.

Our estimate of the half-life of a vehicle is as follows. We record how long the vehicles were not on New Zealand roads by taking the mean difference in features from the MVR: 'vehicle year' and 'first NZ registration year'. We denote this mean difference by YBF for 'years before registration'. Next a local scrappage report [36] can indicate typical ages of vehicles when they are de-registered, which reported an average of 16.7 years and 18.6 years in Wellington and Christchurch, respectively. Thus, we take mean$\{16.7, 18.6\} - YBF$ to be the half-life of vehicles on New Zealand roads. Then, we use the above half-life to take the exponentially weighted moving mean of the 'Emissions intensity from new vehicle registrations' indicator. This constitutes an exponential decay model in terms of the average lifespan of vehicles. There is a precedent for this in the literature, for example [40, Equation (1)]. This gives a variable indicative of the fleet's emissions potential over time.

The product of the fleet emissions potential and the (state highway) traffic volumes variable forms this indicator.

The half-life used above is a rough estimate sufficient only for this proof of concept. In future research we plan to investigate refining our estimate of half-life. This could involve accessing the data in the MVR behind the dashboard mentioned above that provides 'vehicle status'. As well as providing a national view of vehicle activity, this could be updated whenever the MVR is updated, which would be more dynamic that above where the half-life estimate is deduced from a historic scrappage report [36].

A.4. **Weight of coal production.** We extract quarterly values of total coal production from 'Table 1 - Quarterly Tonnes' of the 'Quarterly Coal Supply, Transformation, & Consumption (Tonnes)' publicly available spreadsheet 'Data tables for coal' [3].

Coal is used on a significant scale for energy in the metal, mineral, chemical and manufacturing industries, and construction [33]. Moreover, an increase in coal-fired electricity generation drove an increase in public electricity and heat production emissions between 2018 and 2019 [33].

A.5. **Number of dairy cattle.** We extract annual values for 'Total Dairy Cattle (including Bobby Calves)' from [8]. In 2020, enteric fermentation by dairy cattle made up 35.6% of emissions in the Agriculture sector and has increased by 65.5% since 1990 [35]. This increase is primarily driven by increases in dairy cattle numbers [33, Page 11], which is the purpose of this indicator.

A.6. **Weight of meat produced at export-inspected facilities.** We collate weekly weight statistics of beef and sheep meat production for export reported by the New Zealand Meat Board from certain export-inspected processing facilities dating back to the 2017-18 season. We sum the weight of all types of meat produced across the country. The series is retrofitted back to 1990 with monthly values of 'Total Above Livestock Excluding Game' from [11].

A.7. **Quantity of imported fertilisers.** We extract monthly values for the quantity of imported 'Fertilisers' as in [10]. Use of synthetic nitrogen fertilisers is a driver of emissions trends in Agriculture [33, Page 71]. The use of fertiliser consumption as an indicator has been effective in [12].

A.8. **Food exports.** Food export data is used as a general measure of food production. The indicators 'Food exports for meat', 'Food exports for aquaculture and dairy' and 'Food exports for horticulture' are given by extracting monthly values for 'Live Animals, Meat and Edible Meat Offal', 'Fish, Crustaceans, Molluscs, Dairy Produce and Other Animal Products', and 'Vegetables, Fruit and Prepared Foodstuffs, Beverages and Tobacco' respectively, from [9]. We acknowledge that that this is a price-based indicator, and that price variations as well as production amounts are therefore included. Further development would be to inflation-adjust this indicator to isolate the production component, or replace it entirely with a mass-based food export dataset.

References

[1] [dataset] Energy Market Services. *Current Carbon Intensity API (em6 API)*. https://www.ems.co.nz/em6-api-integration-guide/. Version 1.9. 2022.

[2] [dataset] Ministry for the Environment. *Time series emissions data by category (National inventory report)*. https://environment.govt.nz/publications/new-zealands-greenhouse-gas-inventory-1990-2020/. ME 1635. 2022.

[3] [dataset] Ministry of Business, Innovation and Employment. *Data tables for coal (Energy statistics)*. https://www.mbie.govt.nz/building-and-energy/energy-and-natural-resources/energy-statistics-and-modelling/energy-statistics/coal-statistics/. 2022.

[4] [dataset] New Zealand Meat Board. *Provisional Export Livestock Processing Data*. https://www.nzmeatboard.org/the-industry/production-data/. 2022.

[5] [dataset] New Zealand Transport Agency. *Motor Vehicle Register (Waka Kotahi Open Data)*. https://nzta.govt.nz/resources/new-zealand-motor-vehicle-register-statistics/new-zealand-vehicle-fleet-open-data-sets/#data. 2022.

[6] [dataset] New Zealand Transport Agency. *TMS daily traffic counts API (Waka Kotahi Open Data)*. https://opendata-nzta.opendata.arcgis.com/datasets/tms-daily-traffic-counts-api/api. 2022.

[7] [dataset] New Zealand Transport Agency. *Traffic data booklets and state highway traffic volumes (Waka Kotahi Open Data)*. https://www.nzta.govt.nz/resources/state-highway-traffic-volumes/. 2019.

[8] [dataset] Statistics New Zealand. *AGR001AA (Infoshare)*. http://infoshare.stats.govt.nz/. 2022.

[9] [dataset] Statistics New Zealand. *EXP012AA (Infoshare)*. http://infoshare.stats.govt.nz/. 2022.

[10] [dataset] Statistics New Zealand. *IMP033AA (Infoshare)*. http://infoshare.stats.govt.nz/. 2022.

[11] [dataset] Statistics New Zealand. *LSS025AA (Infoshare)*. http://infoshare.stats.govt.nz/. 2022.

[12] Amaefule, Chukwuemeka, Ibeabuchi, Igwe J., and Shoaga, Akeem. "Determinants of Greenhouse Gas Emissions". In: *European Journal of Sustainable Development Research* 6.4 (June 2022), em0194. DOI: 10.21601/ejosdr/12176.

[13] Bergstra, J and Bengio, Y. "Random Search for Hyper-Parameter Optimization". In: *Journal of Machine Learning Research* 13.10 (2012), pp. 281–305.

[14] Breiman, Leo. "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)". In: *Statistical Science* 16.3 (Aug. 2001). DOI: 10.1214/ss/1009213726.

[15] Calvo Buendia, E. et al. *2019 Refinement to the 2006 IPCC Guidelines for National Greenhouse Gas Inventories*. https://www.ipcc-nggip.iges.or.jp/public/2019rf/index.html. Published: IPCC, Switzerland. 2019.

[16] Chen, Shisheng, Mihara, Kuniaki, and Wen, Jianxiu. "Time series prediction of CO2, TVOC and HCHO based on machine learning at different sampling points". In: *Building and Environment* 146 (Dec. 2018), pp. 238–246. DOI: 10.1016/j.buildenv.2018.09.054.

[17] Debone, Daniela, Leite, Vinicius Pazini, and Miraglia, Simone Georges El Khouri. "Modelling approach for carbon emissions, energy consumption and economic growth: A systematic review". In: *Urban Climate* 37 (May 2021), p. 100849. DOI: 10.1016/j.uclim.2021.100849.

[18] Farrar, Donald E. and Glauber, Robert R. "Multicollinearity in Regression Analysis: The Problem Revisited". In: *The Review of Economics and Statistics* 49.1 (Feb. 1967), p. 92. DOI: 10.2307/1937887.

[19] Filzmoser, Peter, Liebmann, Bettina, and Varmuza, Kurt. "Repeated double cross validation". In: *Journal of Chemometrics* 23.4 (Apr. 2009), pp. 160–171. DOI: 10.1002/cem.1225.

[20] Forster, Piers M. et al. "Current and future global climate impacts resulting from COVID-19". In: *Nature Climate Change* 10.10 (Aug. 2020), pp. 913–919. DOI: 10.1038/s41558-020-0883-0.

[21] Geurts, Pierre, Ernst, Damien, and Wehenkel, Louis. "Extremely randomized trees". In: *Machine Learning* 63.1 (Mar. 2006), pp. 3–42. DOI: 10.1007/s10994-006-6226-1.

[22] Ghorbani, Hamid. "Mahalanobis distance and its application for detecting multivariate outliers". In: *Facta Universitatis, Series: Mathematics and Informatics* (Oct. 2019), p. 583. DOI: `10.2219 0/fumi1903583g`.

[23] González-Sánchez, Mariano and Martín-Ortega, Juan Luis. "Greenhouse Gas Emissions Growth in Europe: A Comparative Analysis of Determinants". In: *Sustainability* 12.3 (Jan. 2020), p. 1012. DOI: `10.3390/su12031012`.

[24] Huang, Cheng et al. "Highly Resolved Dynamic Emissions of Air Pollutants and Greenhouse Gas CO2 during COVID-19 Pandemic in East China". In: *Environmental Science & Technology Letters* 8.10 (Sept. 2021), pp. 853–860. DOI: `10.1021/acs.estlett.1c00600`.

[25] Huang, Yue. "Linear Calibration Methods". In: *Chemometric Methods in Analytical Spectroscopy Technology*. Springer Nature Singapore, 2022, pp. 237–254. DOI: `10.1007/978-981-19-1625-0 _7`.

[26] Iveroth, Sofie Pandis et al. *Nowcasting CO2 emissions*. Nordisk Ministerråd, Aug. 2022. DOI: `10.6027/temanord2022-537`.

[27] Jayaraman, Vaishnavi, Parthasarathy, Saravanan, and Lakshminarayanan, Arun Raj. "Forecasting the Emission of Greenhouse Gases from the Waste using SARIMA Model". In: *2022 6th International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, Apr. 2022. DOI: `10.1109/icoei53556.2022.9777119`.

[28] Kotlar, Ali Mehmandoost, Singh, Jasdeep, and Kumar, Sandeep. "Prediction of greenhouse gas emissions from agricultural fields with and without cover crops". In: *Soil Science Society of America Journal* 86.5 (July 2022), pp. 1227–1240. DOI: `10.1002/saj2.20429`.

[29] Lin, Jinyao et al. "Analyzing the impact of three-dimensional building structure on CO2 emissions based on random forest regression". In: *Energy* 236 (Dec. 2021), p. 121502. DOI: `10.1016 /j.energy.2021.121502`.

[30] Liu, Zhu et al. "Carbon Monitor, a near-real-time daily dataset of global CO2 emission from fossil fuel and cement production". In: *Scientific Data* 7.1 (Nov. 2020). DOI: `10.1038/s41597- 020-00708-7`.

[31] Liu, Zhu et al. "Near-real-time monitoring of global CO2 emissions reveals the effects of the COVID-19 pandemic". In: *Nature Communications* 11.1 (Oct. 2020). DOI: `10.1038/s41467-02 0-18922-7`.

[32] Ministry for the Environment. *Measuring Emissions: A Guide for Organisations: 2020 Detailed Guide. Wellington: Ministry for the Environment*. `https://environment.govt.nz/assets /Publications/Files/Measuring-Emissions-Detailed-Guide-2020.pdf`. Accessed 12 Oct. 2022. 2022.

[33] Ministry for the Environment. *New Zealand's Greenhouse Gas Inventory 1990–2019, National inventory report. Wellington: Ministry for the Environment*. `https://environment.govt.nz/a ssets/Publications/Greenhouse-Gas-Inventory-1990-2019/New-Zealands-Greenhouse- Gas-Inventory-1990-2019-Volume-1-Chapters-1-15.pdf`. Accessed 13 Oct. 2022. 2020.

[34] Ministry for the Environment. *New Zealand's Greenhouse Gas Inventory 1990–2020, National inventory report. Wellington: Ministry for the Environment*. `https://environment.govt.nz/a ssets/publications/GhG-Inventory/New-Zealand-Greenhouse-Gas-Inventory-1990-202 0-Chapters-1-15.pdf`. Accessed 13 Oct. 2022. 2021.

[35] Ministry for the Environment. *New Zealand's Interactive Emissions Tracker*. `https://emissio nstracker.mfe.govt.nz/`. Accessed 8 Nov. 2021. 2022.

[36] Ministry of Transport. *A vehicle scrappage trial for Christchurch and Wellington: May 2009*. `h ttps://www.transport.govt.nz//assets/Uploads/Report/Scrappage-Report-FINAL.pdf`. Accessed 16 Jun. 2021. 2009.

[37] Mutascu, Mihai. "CO2 emissions in the USA: new insights based on ANN approach". In: *Environmental Science and Pollution Research* 29.45 (May 2022), pp. 68332–68356. DOI: `10.1007/s 11356-022-20615-1`.

[38] Ng, Andrew Y. "Preventing "overfitting" of cross-validation data". In: *ICML*. Vol. 97. 1997, pp. 245–253. DOI: `10.5555/645526.657119`.

[39] Oda, Tomohiro et al. "Errors and uncertainties associated with the use of unconventional activity data for estimating CO2 emissions: the case for traffic emissions in Japan". In: *Environmental Research Letters* 16.8 (Aug. 2021), p. 084058. DOI: `10.1088/1748-9326/ac109d`.

[40] Oguchi, Masahiro and Fuse, Masaaki. "Regional and Longitudinal Estimation of Product Lifespan Distribution: A Case Study for Automobiles and a Simplified Estimation Method". In: *Environmental Science & Technology* 49.3 (Jan. 2015), pp. 1738–1743. DOI: 10.1021/es505245 q.

[41] Pedregosa, F. et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[42] Pero, Francesco Del, Delogu, Massimo, and Pierini, Marco. "The effect of lightweighting in automotive LCA perspective: Estimation of mass-induced fuel consumption reduction for gasoline turbocharged vehicles". In: *Journal of Cleaner Production* 154 (June 2017), pp. 566–577. DOI: 10.1016/j.jclepro.2017.04.013.

[43] Quéré, Corinne Le et al. "Temporary reduction in daily global CO2 emissions during the COVID-19 forced confinement". In: *Nature Climate Change* 10.7 (May 2020), pp. 647–653. DOI: 10.103 8/s41558-020-0797-x.

[44] Rogelj, Joeri and Schleussner, Carl-Friedrich. "Reply to Comment on 'Unintentional unfairness when applying new greenhouse gas emissions metrics at country level'". In: *Environmental Research Letters* 16.6 (June 2021), p. 068002. DOI: 10.1088/1748-9326/ac02ec.

[45] Stark, Philip B and Parker, Robert L. "Bounded-variable least-squares: an algorithm and applications". In: *Computational Statistics* 10 (1995), pp. 129–129.

[46] Statistics Netherlands. *Quarterly estimates of greenhouse gas emissions*. https://www.cbs.nl /en-gb/custom/2020/37/quarterly-estimates-of-greenhouse-gas-emissions. Accessed 8 Nov. 2021. 2021.

[47] Statistics New Zealand. *Greenhouse gas emissions (industry and household): March 2021 quarter.* https://www.stats.govt.nz/experimental/greenhouse-gas-emissions-industry-and-ho usehold-march-2021-quarter. Accessed 8 Oct. 2021. 2021.

[48] Ulku, Ilayda and Ulku, Eyup Emre. "Forecasting Greenhouse Gas Emissions Based on Different Machine Learning Algorithms". In: *Lecture Notes in Networks and Systems*. Springer International Publishing, 2022, pp. 109–116. DOI: 10.1007/978-3-031-09176-6_13.

[49] Van Rossum, Guido and Drake, Fred L. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.

[50] Virtanen, Pauli et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[51] Wen, Lei and Yuan, Xiaoyu. "Forecasting CO2 emissions in China's commercial department, through BP neural network based on random forest and PSO". In: *Science of The Total Environment* 718 (May 2020), p. 137194. DOI: 10.1016/j.scitotenv.2020.137194.

[52] Yang, Wenyue and Zhou, Suhong. "Using decision tree analysis to identify the determinants of residents' CO2 emissions from different types of trips: A case study of Guangzhou, China". In: *Journal of Cleaner Production* 277 (Dec. 2020), p. 124071. DOI: 10.1016/j.jclepro.2020.124 071.

[53] Yona, Leehi et al. "Refining national greenhouse gas inventories". In: *Ambio* 49.10 (Jan. 2020), pp. 1581–1586. DOI: 10.1007/s13280-019-01312-9.

[54] Zheng, Bo et al. "Satellite-based estimates of decline and rebound in China's CO2 emissions during COVID-19 pandemic". In: *Science Advances* 6.49 (Dec. 2020). DOI: 10.1126/sciadv.ab d4998.

(M. Jones) *Corresponding author*, DOT LOVES DATA, LEVEL ONE, 80 WILLIS STREET, WELLINGTON CENTRAL, 6011 NEW ZEALAND
Email address: hello@malcolmjones.me

(H. Chorley) DOT LOVES DATA, LEVEL ONE, 80 WILLIS STREET, WELLINGTON CENTRAL, 6011 NEW ZEALAND
Email address: hkchorley@gmail.com

(F. Owen, T. Hilder, H. Trowland, P. Bracewell) DOT LOVES DATA, LEVEL ONE, 80 WILLIS STREET, WELLINGTON CENTRAL, 6011 NEW ZEALAND
Email address: flynn, tamsyn, holly, paul@dotlovesdata.com