Received: 15 January 2021 | Accepted: 20 July 2022

DOI: 10.1111/rssc.12588

ORIGINAL ARTICLE

Journal of the Royal Statistical Society Series C Applied Statistics

Environmental Engel curves: A neural network approach

Tullio Mancini[1] | **Hector Calvo-Pardo**[1,2,3] | **Jose Olmo**[1,4]

[1]Department of Economics, University of Southampton, Southampton, UK

[2]CFS, University of Wisconsin-Madison, Madison, Wisconsin, USA

[3]ILB, Paris, France

[4]Departamento de Análisis Económico, Universidad de Zaragoza, Zaragoza, Spain

Correspondence
Jose Olmo, Department of Economics, University of Southampton, Southampton, UK.
Email: J.B.Olmo@soton.ac.uk

Funding information
Spanish Secretary of Science and Innovation; University of Southampton Presidential Scholarship; Fundación Agencia Aragonesa para la Investigación y el Desarrollo, Grant/Award Number: PID2019-104326GB-I00

Abstract

Environmental Engel curves describe how households' income relates to the pollution associated with the services and goods consumed. This paper estimates these curves with neural networks using the novel dataset constructed in Levinson and O'Brien. We provide further statistical rigor to the empirical analysis by constructing prediction intervals obtained from novel neural network methods such as extra-neural nets and MC dropout. The application of these techniques for five different pollutants allow us to confirm statistically that Environmental Engel curves are upward sloping, have income elasticities smaller than one and shift down, becoming more concave, over time. Importantly, for the last year of the sample, we find an inverted U shape that suggests the existence of a maximum in pollution for medium-to-high levels of household income beyond which pollution flattens or decreases for top income earners.

KEYWORDS

environmental Engel curves, neural networks, prediction uncertainty

1 | INTRODUCTION

Concerns about climate change and the effect of human intervention on global warming have prompted interest in the relationship between household consumption and pollution in

recent years. Much of the early work on environmental economics was aimed to study suitable policies to tax pollution. For example, work by Metcalf (1999) combines the Consumer Expenditure Survey (CEX) with pollution data to study the incidence of a proposed pollution tax. In a related study, Hassett et al. (2009) combine CEX data from several years with pollution data from different industries to show that a carbon tax would be increasingly regressive. Grainger and Kolstad (2010) and Burtraw et al. (2009) use CEX data to show that a carbon tax would be regressive if not offset by lumpsum transfers or reductions in other regressive taxes. More recently, Levinson and O'Brien (2019) explore the concept of Environmental Engel Curves (EEC) and propose a structural approach to estimate the relationship between household income and pollution. This idea extends the concept of Engel curves, see Engel (1895), which study the relationship between households' consumption of particular goods (or services) and households' income.

Environmental Engle curves (EECs) are related to Environmental Kuznets curves that measure the relationship between pollutants and national income, see Grossman and Krueger (1995). However, EECs are structural, representing income expansion paths holding prices constant. Movements along EECs reflect differences in preferences among richer and poorer households within the same social, economics, and regulatory paradigm, holding prices, technologies, and regulation constant. In contrast, shifts in the EECs could be driven by changes in preferences over time, towards a lower pollution content of consumption among US households. Copeland and Taylor (2005) state that the relationship between economic growth and pollution can be described by three separate components: (a) technique (capturing the technologies used for the production and manufacture of goods and services), (b) composition (representing the basket of goods produced by the economy), (c) and scale (quantifies the relation between economic activity and pollution—an increase in economic growth leads to a proportional increase in pollution).

Levinson and O'Brien (2019) compare pollution, income, and consumption for a representative sample of 95,512 US households with annual data over the period 1984 to 2012. These authors construct EECs separately for indirect emissions from each of the five major air pollutants: particulates smaller than 10 microns (PM10), volatile organic compounds (VOCs), nitrogen oxides (NO_x), sulphur dioxide (SO_2) and carbon monoxide (CO), and estimate two versions of each EEC: one based solely on income and one that controls for 18 household characteristics correlated with income, such as education and age. To calculate the pollution emitted by producing the goods and services associated with household expenditures, these authors pair the CEX with emissions intensities calculated from the National Emissions Inventory (NEI).[1] Levinson and O'Brien (2019) find that EECs display three key characteristics. First, these curves are upward sloping, meaning that richer households are responsible for more overall pollution. Second, EECs have income elasticities smaller than one, indicating that although pollution increases with income, top income households' consumption pollution intensity is smaller than for lower income households. And third, EECs shift down and become more concave over time, meaning that for any level of real household income, households in more recent years consume a less polluting mix of goods, the pollution content of which increases with income at a decreasing rate (pollution intensity decreases).

One of the main difficulties associated with the correct study of EECs is the absence of a theoretical framework that describes such relationship. As a consequence, EECs should be

[1]These authors calculate the per dollar emissions intensity of each industry by aggregating industry-level emissions in the 2002 NEI and dividing by the total sales from the 2002 economic and agricultural censuses.

constructed with as few restrictions as possible. For this reason, Levinson and O'Brien (2019) use linear and nonlinear specifications (e.g., cubic polynomials and logarithms) between households' pollution and household-specific information finding similar results across specifications. These authors also consider nonparametric estimates of EECs in their analysis; however, these non-parametric specifications are only considered for the simplest case given by the relationship between household pollution and income. It is well known that non-parametric regression models are not able to accommodate the presence of many covariates, by construction, due to the curse of dimensionality, see Stone (1980).

The aim of the current paper is to investigate the relationship between households' pollution and income using recent state-of-the-art techniques on machine learning (ML). We apply multi-layer neural networks (NNs) to predict non-parametrically the pollution content of household consumption as a function of household income and a large set of covariates. Single and multi-layer NNs are shown to have good theoretical and empirical properties (see, e.g., Cybenko, 1989; Hornik, 1991; Lu et al., 2017) that explain how a sufficiently wide shallow or deep feedforward NN will be able to approximate, accurately, the underlying function, notwithstanding the unknown functional form (Goodfellow et al., 2016). Therefore, we shall employ ReLU feedforward NNs in this work.

Additionally, we propose methods developed in the recent literature on ML to measure the uncertainty around predictions of NNs such as Monte Carlo (MC) dropout, proposed by Gal and Ghahramani (2016), and extra-neural (EN) networks proposed in Mancini et al. (2021). These methods are shown to outperform parametric models such as Hwang and Ding (1997) and non-parametric bootstrap methods, see Tibshirani (1996). A recent contribution by Pomponi et al. (2021) show how the MC dropout reduces significantly the memory requirements associated with ensemble methods. Yet, the predictive intervals obtained from this methodology can be flawed due to the poor approximation provided by the Bayesian variational inference method to the true predictive distribution in some settings.[2] We apply EN networks as a second ML technique to confirm the findings obtained from MC dropout. MC dropout, as shown by Gal and Ghahramani (2016), performs T stochastic forward passes on the same trained NN where the stochasticity is introduced by allowing for dropout not only during training but also prediction phase. The EN nets approach implemented in this paper extends the extra-trees algorithm introduced by Geurts et al. (2006) by estimating T different sub-networks with randomised architectures (each network will have different layer-specific widths) that are independently trained on the same dataset.

By constructing prediction intervals, we are able to attach statistical measures of uncertainty to the pointwise predictions of the model and, hence, add statistical rigor to the empirical findings of Levinson and O'Brien (2019) on the relationship between pollution and income. We construct intervals for the predictions of household pollution along the EEC curve for the reduced model that only considers household income and also for the extended model that incorporates eighteen households' characteristics. As in Levinson and O'Brien (2019), we entertain five different pollutants that capture different dimensions of environmental pollution. The multi-layer NN models that we fit exhibit low mean square prediction errors (MSPE) and, very importantly, accurate coverage probabilities for the associated prediction intervals.

[2]We are grateful to an anonymous referee for raising this issue. Hayashi (2020) derives a lower bound for the difference between the posterior distribution of the model hyperparameters and the approximation provided by the variational Bayesian algorithm and shows that this lower bound can be strictly positive.

Our results replicate to a large extent the empirical findings in Levinson and O'Brien (2019), which is reassuring; however, we obtain two findings that contrast with this seminal study and help to highlight the importance of considering ML techniques in this context. First, we find an increase in the concavity of the EEC when we increase the number of covariates with respect to the model that only contains household income (and income squared). In contrast, the concavity of the EEC decreases in Levinson and O'Brien (2019) as we increase the number of covariates. Related to this is our second insight; for 2012, we find that the pollution intensity of US households reaches a peak at medium-to-high income levels, and then decreases for top income earners. This interesting finding is only observed in Levinson and O'Brien (2019) for the simple parametric model with income and income squared, but we find robust evidence of this phenomenon when also considering the model with eighteen household covariates, for the five pollutants studied in Levinson and O'Brien (2019).

The rest of the paper is organised as follows: Section 2 reports the definitions and notations used in the paper. Section 3 introduces the EN net approach recently developed in Mancini et al. (2021) and the popular MC dropout approach (see Gal and Ghahramani, 2016). Section 4 reports and discusses the empirical results. Section 5 concludes. Figures are reported at the end of the document.

2 | NN BASICS

The present section provides the notation and definitions used in the remainder of the paper. It will first define Rectified Linear Unit (ReLU) activation functions and multi-layer (sequential) feedforward NNs with emphasis on univariate regression tasks and the definition of dropout.

2.1 | Definition and notations

Let $y_i \in \mathbb{R}$ for $i = 1, \dots, n$ denote the outcome variable and $\mathbf{x}_i = (x_{1i}, \dots, x_{di})$ a set of input variables (covariates) used to predict y_i. A general specification to describe the relationship between both sets of variables is

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \tag{1}$$

with $f(\mathbf{x})$ a real-valued function, and ϵ an error term that satisfies $\mathbb{E}[\epsilon_i|\mathbf{x}_i] = 0$. In standard regression settings the question of interest is to approximate the unknown function $f(\mathbf{x})$. It is well known that if the function $f(\mathbf{x})$ is linear on $\mathbf{x}$, under standard regularity conditions on the error term, ordinary least square (OLS) regression methods provide unbiassed, consistent and efficient estimators of the model coefficients. However, many empirical problems are characterised by non-linear relationships between the variables of interest. The presence of a large number of covariates also compromises the good theoretical properties of OLS methods in many regression settings. In this particular case, Levinson and O'Brien (2019) explain how there is no theory that defines the form of the income-pollution relationship and thus, the construction and analysis of the EECs should be conducted with as few restrictions as possible. Non-parametric kernel regression models are a possible solution for the simplified model that only considers the relationship between household income and pollution, but it is not a feasible option when we also consider the presence of additional covariates capturing households characteristics.

Recent advances in ML have shown that NN methods provide accurate predictions without requiring specific knowledge of the underlying data generating process and thus, they are not

affected by parametric model misspecification. In this setting, single and multi-layer NNs provide a powerful tool for the construction of EECs and thus, the present paper considers $f(\mathbf{x})$ to be in the class of fully connected feedforward neural networks (or multi-layer perceptrons, MLP) with ReLU activation functions.[3]

A ReLU activation function can be defined as follows. Let $\boldsymbol{\theta}(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^d$, with

$$\boldsymbol{\theta}(\mathbf{x}) = (\max\{0, x_1\}, \max\{0, x_2\}, \ \dots \ , \max\{0, x_d\}), \tag{2}$$

where d denotes the number of covariates (input dimension). Alternatively, the ReLU activation function can be expressed as $\boldsymbol{\theta}(\mathbf{x}) = \mathbb{I}(\mathbf{x} > 0) \cdot \mathbf{x}$, with $\mathbb{I}(\mathbf{x} > 0)$ the indicator function.

Having defined the ReLU activation function, it is now possible to provide a definition of a multi-layer ReLU NN. For any two natural numbers $d, \ n_1 \in \mathbb{N}$, which are called input and output dimension, respectively, a $\mathbb{R}^d \to \mathbb{R}^{n_1}$ ReLU neural network is given by specifying a natural number $N \in \mathbb{N}$, a sequence of N natural numbers $Z_1, Z_2, \ \dots \ , Z_N$, and a set of $N + 1$ affine transformations $\mathbf{T}_1 : \mathbb{R}^d \to \mathbb{R}^{Z_1}, \mathbf{T}_i : \mathbb{R}^{Z_{i-1}} \to \mathbb{R}^{Z_i}$, for $i = 2, \ \dots \ , N$, and $\mathbf{T}_{N+1} : \mathbb{R}^{Z_N} \to \mathbb{R}^{n_1}$. Such NN is called a $(N + 1)$-layer ReLU NN, and is said to have N hidden layers. The function $f : \mathbb{R}^d \to \mathbb{R}^{n_1}$ is the output of this ReLU NN that is constructed as

$$f(\mathbf{x}; \boldsymbol{\omega}) = \mathbf{T}_{N+1} \circ \boldsymbol{\theta} \circ \mathbf{T}_N \circ \cdots \circ \mathbf{T}_2 \circ \boldsymbol{\theta} \circ \mathbf{T}_1, \tag{3}$$

with $\mathbf{T}_n = \mathbf{W}^n \mathbf{h}_{n-1} + \mathbf{b}_n$, where—for $N = 1 - \mathbf{W}^n \in \mathbb{R}^{Z_1 \times d}$; $\mathbf{h}_0 \equiv \mathbf{x}$, with $\mathbf{x} \in \mathbb{R}^{d \times 1}$ the input layer, and $\mathbf{b}_n \in \mathbb{R}^{Z_1}$ is an intercept or bias vector. For $N \neq 1$, $\mathbf{W}^n \in \mathbb{R}^{Z_n \times Z_{n-1}}$ is a matrix with the deterministic weights determining the transmission of information across layers; $\mathbf{h}_{n-1} \in \mathbb{R}^{Z_{n-1}}$ is a vector defined as $\mathbf{h}_{n-1} = \boldsymbol{\theta}(\mathbf{T}_{n-1})$, and $\mathbf{b}_n \in \mathbb{R}^{Z_n}$. The function $\boldsymbol{\theta}$ is a ReLU activation function defined as $\boldsymbol{\theta}(\mathbf{T}_{n-1}) = \max\{0, \mathbf{T}_{n-1}\}$ and $\boldsymbol{\omega} = \{\boldsymbol{W}^n, \mathbf{b}_n\}_{n=1}^N$ collects the set of estimable features of the model. The *depth* of a ReLU NN is defined as $N + 1$. The width of the nth hidden layer is Z_n, and the *width* of a ReLU NN is $\max\{Z_1, \ \dots \ , Z_N\}$. The *size* of the ReLU NN is $Z_{\text{tot}} = Z_1 + Z_2 + \cdots + Z_N$, that corresponds to the total number of nodes in the NN architecture. The number of active weights (different from zero) in the nth hidden layer is $w_n = (Z_n \times Z_{n-1}) + Z_n$. The *number of active weights* in a fully connected ReLU NN is $w_1 + w_2 + \cdots + w_N$. The same definition applies to single-layered networks by imposing $N = n = 1$.

In practice, there is an approximation error due to replacing $f(\mathbf{x})$ by $f(\mathbf{x}; \boldsymbol{\omega})$ in model (1), where $f(\mathbf{x}; \boldsymbol{\omega})$ denotes a feasible version of the multi-layer NN model that can be estimated from the data.[4] The model that we consider in practice is

$$y_i = f(\mathbf{x}_i; \boldsymbol{\omega}) + u_i, \tag{4}$$

where $u_i = \varepsilon_i + f(\mathbf{x}_i) - f(\mathbf{x}_i; \boldsymbol{\omega})$ is the sum of the idiosyncratic error ε_i and an approximation error $f(\mathbf{x}_i) - f(\mathbf{x}_i; \boldsymbol{\omega})$ that is negligible for suitable architectures of the NN and sufficiently large sample sizes.

The construction of prediction intervals around the pointwise predictions of NN models has most recently been object of important research in ML applications. The possibility of

[3]Other prominent examples in the ML literature include support vector machines (SVMs), boosting algorithms (e.g., Adaboost), decision trees (and their generalisation to random forests and extremely randomised trees), and non-parametric regressions in the spirit of nearest neighbors and local kernel smoothing.

[4]The feasible NN is defined by a truncation of the true ReLU multi-layer NN that approximates arbitrarily well the unknown function $f(\mathbf{x})$.

constructing prediction intervals allows us to measure the uncertainty around the model predictions. The concept of MC dropout is central to this novel literature on prediction intervals for NN models, see Srivastava et al. (2014). Before discussing the construction of prediction intervals, we elaborate on the concept of dropout in NN models.

Training with dropout (*dropout training*) implies that for each iteration of the learning algorithm different random sub-networks (or *thinned* networks) are trained. Let h_{zn} denote the elements of the vector $\mathbf{h}_n$ for a given node $z = 1, \dots, Z_n$ in layer $n = 1, \dots, N$. Srivastava et al. (2014) develop a dropout methodology that is applied to each function h_{zn} to obtain a transformed variable $\overline{h}_{zn}$. This variable is obtained by pre-multiplying h_{zn} by a random variable r_{zn} with distribution function $F(r_{zn})$, such that $\overline{h}_{zn} = r_{zn} \cdot h_{zn}$, for all (z, n), prior to being fed forward to the activation function in the next layer, h_{zn+1}, for all $z = 1, \dots, Z_{n+1}$. For any layer n, $\mathbf{r}_n = [r_{1n}, \dots, r_{Z_n n}] \in \mathbb{R}^{Z_n}$ denotes a vector of independent random variables. In the empirical application, we consider only the Bernoulli probability distribution $F(r_{zn})$, where each r_{zn} has probability p of being 1 (and $q = 1 - p$ of being 0). The vector $\mathbf{r}_n$ is then sampled and multiplied element-wise with the outputs of that layer, h_{zn}, to create the thinned outputs, $\overline{h}_{zn}$, which are then used as input to the next layer, h_{zn+1}. When this process is applied at each layer $n = 1, \dots, N$, this amounts to sampling a sub-network from a larger network at each forward pass (or gradient step). At test time, the weights are scaled down as $\overline{\mathbf{W}}^n = p\mathbf{W}^n$, $n = 1, \dots, N$, returning a deterministic output[5]. We then identify $\mathbf{r}^\star = [\mathbf{r}_1, \dots, \mathbf{r}_N]$ as the collection of independent random variables applied to a feedforward NN of depth $N + 1$.

3 | PREDICTION INTERVALS FOR MULTI-LAYER NNS

The construction of prediction intervals around the pointwise predictions of multi-layer NNs has most recently been object of important research in ML applications. The possibility of constructing prediction intervals allows us to measure the uncertainty around the model predictions. Prediction intervals for both single and multi-layer neural networks are derived from the predictive distribution of the model output. Hwang and Ding (1997) are the first authors to propose an asymptotic prediction interval for single-layer NNs. Despite the theoretical appeal of this approach its implementation in large dimensions—and under the absence of a parametric setting—has important limitations associated with the correct computation of the Jacobian matrix of the model specification, see, for example, Tibshirani (1996) and Devieaux et al. (1998).

Recent work on NN models introduces uncertainty through bootstrap resampling techniques and MC simulation methods enabling the construction of prediction intervals for the outputs of multi-layer NNs. When focusing on the first sub-group, pairs and residual bootstrapping can be regarded as the methodologies most adopted by practitioners (see Dipu Kabir et al., 2018; Tibshirani, 1996; and Heskes, 1997 for reviews on the topic). Despite its importance in recent empirical work, the relevant literature identifies several limitations associated with bootstrapping methods: (i) Lee et al. (2015) show how re-sampling with replacement reduces the number of unique observations used to train the model by 37%; (ii) Lee et al. (2015) and Lakshminarayanan et al. (2017) show, empirically, how data re-sampling in ensembles of NNs deteriorates not only the prediction accuracy but also the definition of the predictive uncertainty of the ensemble itself;

[5]In practice, an inverted dropout methodology is applied when implementing this methodology in Keras for RStudio. In this case, instead of scaling-down the weights at test time, the weights are scaled-up during train time as $\overline{\mathbf{W}}^n = (1/p)\mathbf{W}^n$, $n = 1, \dots, N$. At test time, a single deterministic forward pass on the unscaled weights $\mathbf{W}^n$ is performed.

(iii) El Karoui and Purdom (2018) show that both pairs and residual bootstrapping suffer from several problems when applied in high-dimensional linear regression problems. In particular, the residual bootstrapping tends to give under-conservative estimates of the uncertainty, while the pairs bootstrapping provides over-conservative estimates.

Gal and Ghahramani (2016) propose an alternative approach for the approximation of the predictive distribution of NNs called MC dropout. The concept of MC dropout is central to this novel literature on prediction intervals for NN models, see Srivastava et al. (2014). An alternative to construct prediction intervals in a NN framework, recently proposed in Mancini et al. (2021), explores the results for randomised trees and derives valid confidence intervals of the model predictions in finite samples. In this work, we focus on the latter two methodologies that are reviewed in the following subsections. We start with the EN network approach as it may be less known than MC dropout.

3.1 | EN network

The EN network approach can be interpreted as an ensemble predictor for NN models. This methodology, formally introduced in Mancini et al. (2021), is based on the extremely randomised trees approach proposed by Geurts et al. (2006) for random forests.

For notation purposes, we will identify the fixed Bernoulli mask as $\bar{\mathbf{r}}^{\star}$ as opposed to $\mathbf{r}^{\star}$ usually employed for dropout training. In this setting, T sets of vectors $\{\bar{\mathbf{r}}^{\star(t)}\}_{t=1}^{T}$ are sampled from a Bernoulli distribution prior to training and are kept constant during both train and test phases. This approach reduces to train and independently fit T random sub-networks on the same dataset. In this setting, generating the predictive distribution is similar, in spirit, to an ensemble approach that trains different sub-neural networks on the same dataset. We consider T fitted sub-networks defined as $f_t(\mathbf{x}_i; \widehat{\boldsymbol{\omega}}^{(t)})$, with $t = 1, \dots, T$. We use f_t to note that each prediction belongs to a potentially different NN model. Given the following ensemble, let $\bar{f}_{\text{EN}}(\mathbf{x}_i)$ denote the ensemble predictor obtained from the EN network approach that is constructed as

$$\bar{f}_{\text{EN}}(\mathbf{x}_i) = \frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{x}_i; \widehat{\boldsymbol{\omega}}^{(t)}), \text{ for } i = 1, \dots, n. \tag{5}$$

where $\widehat{\boldsymbol{\omega}}^{(t)}$ denotes the parameter estimates obtained from fitting each sub-network independently. Mancini et al. (2021) show that the MSPE of the above ensemble predictor can be expressed as

$$\text{MSPE}(\bar{f}_{\text{EN}}(\mathbf{x}_i)) = \mu_i^2 + \frac{1}{T}\sigma^2_{\widehat{\boldsymbol{\omega}}}(\mathbf{x}_i) + \frac{T-1}{T}c_i. \tag{6}$$

This expression extends Zhou (2012) by showing that the MSPE of the ensembler (5) depends on the variance of the individual predictor models, their covariance, c_i, and the approximation bias μ_i. The smaller the covariance, the smaller the generalisation error of the ensemble. In contrast, if the different predictors are perfectly correlated (as for the MC dropout) we know that $c_i = \sigma^2_{\widehat{\boldsymbol{\omega}}}(\mathbf{x}_i)$ and thus MSPE $(\bar{f}_{\text{EN}}(\mathbf{x}_i)) = \sigma^2_{\widehat{\boldsymbol{\omega}}}(\mathbf{x}_i)$ - effectively reducing to zero the effect of ensembling. Similarly, the MSPE is minimised when the errors are perfectly uncorrelated and thus when $c_i = 0$. This result has important implications when analysing the epistemic uncertainty of an EN network. Given the zero correlation between the predictions of the sub-networks (see Mancini

et al., 2021 for additional experimental results), then as $T \to \infty$, the MSPE($\bar{f}_{\text{EN}}(\mathbf{x}_i)$) converges to zero, assuming that the approximation bias is negligible. Therefore, a suitable prediction interval is

$$\bar{f}_{\text{EN}}(\mathbf{x}_i) \pm z_{1-\alpha/2}\left(\frac{\hat{\sigma}^2_{\hat{\boldsymbol{\omega}}}(\mathbf{x}_i)}{T} + \hat{\sigma}^2_{\epsilon}\right)^{1/2}, \tag{7}$$

with $\hat{\sigma}^2_{\hat{\boldsymbol{\omega}}}(\mathbf{x}_i) = \frac{1}{T}\sum_{t=1}^{T}(f_t(\mathbf{x}_i;\hat{\boldsymbol{\omega}}^{(t)}) - \bar{f}_{\text{EN}}(\mathbf{x}_i))^2$ and $\hat{\sigma}^2_{\epsilon} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{f}_{\text{EN}}(\mathbf{x}_i)\right)^2$, where n is the size of the test sample.[6]

As explained in Zhou (2012), the covariance term in Equation (6) captures the diversity existing among the T different sub-networks identifying the EN network. The aim of the EN network approach is to construct individual predictors that are mutually independent such that the prediction interval (7) is valid.

In order to generate $\{f_t(\mathbf{x};\hat{\boldsymbol{\omega}}^{(t)})\}_{t=1}^{T}$, we sample T vectors $\{\bar{\mathbf{r}}^{\star(t)}\}_{t=1}^{T}$ prior to training. Each fixed Bernoulli mask is applied independently to the original network returning T independent sub-networks of size $Z^{(t)}_{\text{extra net}} \leq Z_{\text{tot}}$, where Z_{tot} denotes the total number of nodes in the NN. Each sub-network is then trained independently on $\mathbf{x}^{\star}$, and T deterministic forward passes are performed at test phase. The procedure reported in Algorithm 1 shows that an EN network is an ensemble of T NNs with randomised weights and structures and no data re-sampling. By randomising not only the weights of the T sub-networks but also their structure, and by fitting the networks on the entire training set $\{y_i;\mathbf{x}_i\}_{i=1}^{M}$, this method outperforms the bootstrap approach in terms of both out-of-sample prediction accuracy (Lee et al., 2015) and uncertainty quantification (Lakshminarayanan et al., 2017). See also the comparison study in Mancini et al. (2021) with respect to the bootstrapping ensemble approach and MC dropout methods.

The algorithm to implement this approach is as follows:

3.2 | MC dropout

MC dropout was originally developed for Bayesian NNs (Denker & LeCun, 1991) and subsequently extended beyond the Bayesian framework by Cortes-Cirano and Bender (2019), among other authors. This approach introduces randomness into the NN prediction by implementing dropout not only during training but also during testing.

Gal and Ghahramani (2016) propose a new theoretical framework which uses dropout in NNs as approximate Bayesian inference for deep Gaussian processes. In this sub-section we adopt this methodology outside Bayesian NNs and illustrate how to construct prediction intervals for the output y_i. The literature focusing on Bayesian deep NNs concentrates on correctly approximating the posterior probability distribution of the output of the NNs, which is often intractable. More specifically, let $p(\hat{y}\,|\mathbf{x},\mathbf{X},\mathbf{Y})$ denote the distribution of the predictive output $\hat{y}$ conditional on the set of observations $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\mathbf{Y} = \{y_1, \ldots, y_n\}$. The predictive probability distribution of the NN model is

$$p(\hat{y}\,|\mathbf{x},\mathbf{X},\mathbf{Y}) = \int_{\boldsymbol{\Omega}} p(\hat{y}\,|\mathbf{x},\boldsymbol{\omega})p(\boldsymbol{\omega}\mid\mathbf{X},\mathbf{Y})\mathrm{d}\boldsymbol{\omega}, \tag{11}$$

[6]Note that for obtaining a consistent estimator of $\hat{\sigma}^2_{\epsilon}$ we have imposed homoscedasticity of the error terms ϵ_i over the test sample.

Algorithm 1. EN networks

INPUT: Training Data $\{x_i^{\text{tr}} \equiv (\mathbf{x}_i, y_i)\}_{i=1}^{M}$

OUTPUT: Prediction Interval $\widehat{f}(\mathbf{x}; \boldsymbol{\omega})$.

procedure T LEARNERS

Define depth and width of *original* NN.

while (t < T) **do**

Generate a Bernoulli mask $\bar{\mathbf{r}}^{\star}$ prior to training.

Apply Bernoulli mask $\bar{\mathbf{r}}^{\star}$ to the *original* NN.

Train random thinned network on $\mathbf{x}^{\text{tr}}$ with random initialisation of $\{\mathbf{W}_0^n\}_{n=1}^{N}$

Trained thinned network → Deterministic forward pass on test data.

Store $f_t(\mathbf{x}_i; \widehat{\boldsymbol{\omega}}^{(t)})$.

end while

Compute the ensemble estimate:

$$\bar{f}_{\text{EN}}(\mathbf{x}_i) = \frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{x}_i; \widehat{\boldsymbol{\omega}}^{(t)}). \tag{8}$$

Compute the epistemic and aleatoric variance:

$$\begin{cases} \widehat{\sigma}^2_{\widehat{\boldsymbol{\omega}}}(\mathbf{x}_i) = \frac{1}{T}\sum_{t=1}^{T}[f_t(\mathbf{x}_i; \widehat{\boldsymbol{\omega}}^{(t)}) - \bar{f}_{\text{EN}}(\mathbf{x}_i)]^2 \\ \widehat{\sigma}^2_{\epsilon} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{f}_{\text{EN}}(\mathbf{x}_i)\right)^2 \end{cases}. \tag{9}$$

Define Prediction interval:

$$\bar{f}_{\text{EN}}(\mathbf{x}_i) \pm z_{1-\alpha/2}\widehat{\sigma}_e, \tag{10}$$

with $\widehat{\sigma}_e = \left(\frac{\widehat{\sigma}^2_{\widehat{\boldsymbol{\omega}}}(\mathbf{x}_i)}{T} + \widehat{\sigma}^2_{\epsilon}\right)^{1/2}$.

return Prediction interval (10)

end procedure

with $p(\widehat{y}\,|\mathbf{x}, \boldsymbol{\omega})$ the likelihood function of the observations, and $\boldsymbol{\omega} \in \boldsymbol{\Omega}$ where $\boldsymbol{\Omega}$ denotes the parameter space. The posterior probability distribution $p(\boldsymbol{\omega} \mid \mathbf{X}, \mathbf{Y})$ is intractable.

Gal and Ghahramani (2016) propose NN dropout to approximate this distribution. More formally, under model dropout, we consider a distribution function $q(\boldsymbol{\omega})$ that follows a Bernoulli distribution, Ber(p). The above predictive distribution in this Bayesian NN setting can be approximated by

$$p(\widehat{y}\,|\mathbf{x}, \mathbf{X}, \mathbf{Y}) = \int_{\Omega} p(\widehat{y}\,|\mathbf{x}, \boldsymbol{\omega})q(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}. \tag{12}$$

In practice this predictive distribution can be approximated using MC methods. Thus, by sampling T sets of vectors from the Bernoulli distribution $\{\mathbf{r}^{\star(t)}\}_{t=1}^{T}$, one can approximate the above predictive distribution from the random sample $\widehat{y}(\mathbf{x}_i; \widehat{\boldsymbol{\omega}}^{(t)})$, for $i = 1, \ldots, n$, where $\widehat{\boldsymbol{\omega}}^{(t)} =$

$\{\widehat{\mathbf{W}}^{1(t)}, \ldots, \widehat{\mathbf{W}}^{N(t)}, \widehat{b}_1^{(t)}, \ldots, \widehat{b}_N^{(t)}\}$ denotes the sequence of weights associated to the different nodes and layers of the NN and the associated bias parameters for a given pass t for $t = 1, \ldots, T$.

Using this MC dropout technique, Gal and Ghahramani (2016) propose the first moment from the MC predicted outputs as the model prediction:

$$\bar{f}_{\mathrm{MC}}(\mathbf{x}_i) = \frac{1}{T}\sum_{t=1}^{T}\widehat{y}(\mathbf{x}_i; \widehat{\boldsymbol{\omega}}^{(t)}), \text{ for } i = 1, \ldots, n. \tag{13}$$

These authors show that, in practice, this is equivalent to performing T stochastic forward passes through the network and averaging the results. This result has been presented in the literature before as model averaging. Srivastava et al. (2014) have reasoned empirically that MC dropout can be approximated by averaging the weights of the network (multiplying each weight $\mathbf{W}^n$ by p at test time, and referred to as standard dropout).

Importantly, the model parameters $\boldsymbol{\omega}$ are fixed across random samples implying that the cross-correlation between the predictions $\widehat{y}(\mathbf{x}_i; \widehat{\boldsymbol{\omega}}^{(t)})$ and $\widehat{y}(\mathbf{x}_i; \widehat{\boldsymbol{\omega}}^{(t')})$ for $t, t' = 1, \ldots, T$ is perfect. Then, the predictive variance is defined as

$$\sigma^2_{\mathrm{MC}} = \widehat{\sigma}^2_{\epsilon} + \frac{1}{T^2}\sum_{t=1}^{T}\sum_{t'=1}^{T} E\left[\left(\widehat{y}(\mathbf{x}_i; \widehat{\boldsymbol{\omega}}^{(t)}) - E[\widehat{y}(\mathbf{x}_i; \widehat{\boldsymbol{\omega}}^{(t)})]\right)\left(\widehat{y}(\mathbf{x}_i; \widehat{\boldsymbol{\omega}}^{(t')}) - E[\widehat{y}(\mathbf{x}_i; \widehat{\boldsymbol{\omega}}^{(t')})]\right)\right], \tag{14}$$

The first component on the right-hand side expression of (14) captures the aleatoric uncertainty whereas the second term captures the epistemic uncertainty associated to parameter estimation. The second term includes the estimation of the variance and covariance terms between the different random samples obtained from using dropout. Thus, under the assumption that the approximation error is negligible, the above predictive variance can be estimated as

$$\widehat{\sigma}^2_{\mathrm{MC}} = \widehat{\sigma}^2_{\epsilon} + \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{y}(\mathbf{x}_i; \widehat{\boldsymbol{\omega}}^{(t)}) - \bar{f}_{\mathrm{MC}}(\mathbf{x}_i)\right)^2, \tag{15}$$

with $\widehat{\sigma}^2_{\epsilon} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{f}_{\mathrm{MC}}(\mathbf{x}_i)\right)^2$ a consistent estimator of σ^2_{ϵ} under homoscedasticity of the error term, see also Gal and Ghahramani (2016) and Kendall and Gal (2017). A suitable prediction interval for y_i under the assumption that $p(\widehat{y} \,|\mathbf{x}, \boldsymbol{\omega})$ is normally distributed is

$$\bar{f}_{\mathrm{MC}}(\mathbf{x}_i) \pm z_{1-\alpha/2}\widehat{\sigma}_{\mathrm{MC}}. \tag{16}$$

To further understand the concept of stochastic forward passes, we consider the following definition of a generic multi-layer NN:

$$f(\mathbf{x}_i; \boldsymbol{\omega}) = \mathbf{W}_N\theta(\cdots\theta(\mathbf{W}_2\theta(\mathbf{W}_1\mathbf{x}_i + \boldsymbol{b}_1) + \boldsymbol{b}_2) + \cdots) + \boldsymbol{b}_N. \tag{17}$$

Given the trained NN in (17), the predictions (test time) are obtained by the following matrix multiplication:

$$\widehat{y}_i = \widehat{\mathbf{W}}_N\widehat{\theta}(\cdots\widehat{\theta}(\widehat{\mathbf{W}}_2\widehat{\theta}(\widehat{\mathbf{W}}_1\mathbf{x}_i + \widehat{\boldsymbol{b}}_1) + \widehat{\boldsymbol{b}}_2) + \cdots) + \widehat{\boldsymbol{b}}_N. \tag{18}$$

Performing T stochastic forward passes is equivalent to performing T stochastic matrix multiplications where the stochasticity is introduced by the T Bernoulli masks applied to the network weights at test time. This implies that by using the Bernoulli masks also at test time, the predictions obtained from MC dropout can be formulated as follows:

$$\widehat{y}(\mathbf{x}_i;\widehat{\boldsymbol{\omega}}^{(t)}) = \widehat{\mathbf{W}}_N^{(t)}\widehat{\boldsymbol{\theta}}(\cdots\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{W}}_2^{(t)}\widehat{\boldsymbol{\theta}}(\widehat{\mathbf{W}}_1^{(t)}\mathbf{x}_i + \widehat{\boldsymbol{b}}_1^{(t)}) + \widehat{\boldsymbol{b}}_2^{(t)}) + \cdots) + \widehat{\boldsymbol{b}}_N^{(t)}. \tag{19}$$

In other words, by implementing MC dropout, the deterministic prediction defined in Equation (18), is replaced by a stochastic operation (defined in Equation 19) where the level of stochasticity is dictated by the sets of Bernoulli mask $\{\mathbf{r}^{\star(t)}\}_{t=1}^{T}$ randomly sampled.

Additionally, Equation (18) allows understanding also the difference between the MC dropout and the extra NN approach conveyed by Equations (7) and (13). More precisely, by $\widehat{y}(\mathbf{x}_i;\boldsymbol{\omega}^{(t)})$ we indicate the predictions obtained from the *same* predictive model $f(\mathbf{x}_i;\boldsymbol{\omega})$ when the t random subsets of weights $\boldsymbol{\omega}^{(t)}$—identified via the random Bernoulli mask $\mathbf{r}^{\star(t)}$—are adopted; by $f_t(\mathbf{x}_i;\boldsymbol{\omega})$ we indicate the t predictive models. Thus, the computational and memory requirements needed for the implementation of the MC dropout are sensibly lower than the ones required from the extra NN algorithm[7] (Algorithm 2).

Algorithm 2. MC Dropout

INPUT: Training Data $\{x_i^{\llcorner} \equiv (\mathbf{x}_i, y_i)\}_{i=1}^{M}$

OUTPUT: Prediction Interval $\widehat{f}(\mathbf{x};\boldsymbol{\omega})$.

1: **procedure** T STOCHASTIC FORWARD PASSES

2:

3: Define depth and width of *original* NN.

4: Train NN with dropout on $\mathbf{x}^{\llcorner}$ without scaling up the weights.

5: **while** (t < T) **do**

6: Stochastic forward pass from the trained NN on test data.

7: Store $\widehat{y}(\mathbf{x}_i;\widehat{\boldsymbol{\omega}}^{(t)})$

8: **end while**

9: Compute the ensemble estimate:

$$\bar{f}_{\text{MC}}(\mathbf{x}_i) = \frac{1}{T}\sum_{t=1}^{T}\widehat{y}(\mathbf{x}_i;\widehat{\boldsymbol{\omega}}^{(t)}),\ \text{for } i = 1,\dots,n. \tag{20}$$

10: Compute the epistemic and aleatoric variance:

$$\begin{cases}\widehat{\sigma}^2_{\text{MC}} = \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{y}(\mathbf{x}_i;\widehat{\boldsymbol{\omega}}^{(t)}) - \bar{f}_{\text{MC}}(\mathbf{x}_i)\right)^2 \\ \widehat{\sigma}^2_{\epsilon} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{f}_{\text{MC}}(\mathbf{x}_i)\right)^2\end{cases}. \tag{21}$$

11: Define Prediction Interval:

$$\bar{f}_{\text{MC}}(\mathbf{x}_i) \pm z_{1-\alpha/2}\widehat{\sigma}_{\text{MC}}. \tag{22}$$

with $\widehat{\sigma}_e = \left(\widehat{\sigma}^2_{\text{MC}} + \widehat{\sigma}^2_{\epsilon}\right)^{1/2}$.

return Prediction interval (22)

12: **end procedure**

[7]Nonetheless, the memory requirements associated with the extra neural network algorithm are lower than a normal ensemble of T NNs as also shown in Pomponi et al. (2021).

4 | EMPIRICAL RESULTS

Levinson and O'Brien (2019) find that EECs display three key characteristics: EECs are upward sloping, have income elasticities smaller than one, shifting down and becoming more concave over time. These features of the relationship between the variables can be quantified and assessed through the construction of prediction intervals for a given coverage probability. Thus the hypothesis that the relationship is increasing could be tested by assessing if the first derivative of the predicted function modelling the relationship between pollution and household income is strictly positive. Similarly, the concavity of the relationship could be tested by assessing if the second derivative of the predicted functional form is negative. Once the concavity of the function is not rejected, the third hypothesis given by an income elasticity smaller than one can be tested by assessing if the slope of the functional form relating household pollution and income is less than one uniformly over the relevant domain.[8]

Following Levinson and O'Brien (2019) we report results for the years 1984 and 2012 and consider five different pollutants: particulates smaller than 10 microns (PM10), VOCs, NO_x, SO_2, and CO.[9] Table 1 in Levinson and O'Brien (2019) presents the relevant summary statistics for the different variables observed for the year 1984 and 2012. The table reports the average values and the standard errors (when possible) of the variables comprising the dataset. The number of observations adopted for the study is 3184 for the year 1984 and 3538 for the year 2012. These authors construct two types of EECs: one using income and squared income as only covariates and an extended version of the model in which the set of covariates is expanded to incorporate other available households' characteristics. In the latter case, these authors consider $d = 18$ regressors to predict the pollution content of consumption. Table 2 in Levinson and O'Brien (2019) lists the full set of socio-economic, demographic and spatial covariates used.

Due to differences in dimensionality across problems (using income only or adding household covariates), in the first case, a set of candidate single-layer NNs with $Z_{\text{tot}} = [5, 10, 20]$ is considered. In the second case, we consider a multi-layer NN with architecture (width and depth) obtained using the constrained optimisation approach in Calvo-Pardo et al. (2021), that maximises the minimum number (lower bound) of linear regions approximated by a ReLU NN (the interested reader is referred to Montufar et al., 2014). As a result, an optimal allocation of hidden nodes obtains, both within (width) and across (depth) hidden layers, for fully connected feedforward NNs. The NN size considered at the onset depends on the complexity of the functional form to be approximated. The novel procedure reduces the computational requirements associated with the identification of the neural network structure, since it only requires cross-validating the size of the optimised NN structure—instead of cross-validating both size and hidden nodes allocation.

This procedure is adopted to identify the structure of the *original* NN, from which either T random sub-networks are obtained (EN network approach), or which is trained with dropout (MC dropout approach)—not only at train but also during validation. Thus, NNs with different number of total nodes $Z_{\text{tot}} = [76, 90, 150, 200, 250, 392, 446, 500]$ are tuned and optimised to obtain optimal architectures.[10]

[8]The reader should note that this definition of elasticity is different from the general definition of the elasticity of Y with respect to X that is given by $E_X^Y = \frac{\%\text{ change in } Y}{\%\text{ change in } X}$, which reduces to $E_X^Y = \frac{dY}{dX}\frac{X}{Y}$ for infinitesimal changes and differentiable variables.

[9]Results for the years 1985–2011 are available from the authors upon request.

[10]The choice of the total number of nodes is the result of a numerical optimisation exercise. For each exercise, we consider an interval [min, max] with realistic values for the total number of nodes, and construct a grid of values spanning the interval. Our procedure retains the value that minimises the MSPE of the NN out of sample.

The hyper-space defined by the different optimisation algorithms, weights initialisers, learning rates, number of epochs, and drop-out rates are defined with no distinction between single and multi-layer NNs. In particular, the learning rates 0.001, 0.003, 0.00, and 0.01 for the Adam optimiser ($\beta_1 = 0.9$; $\beta_2 = 0.999$), and for the RMSProp optimiser with $\rho = 0.9$ are tuned. When the Adam and RMSProp optimiser are analysed, the He normal initialiser and the Xavier uniform (default in Keras) initialiser are implemented. The former draws samples from a truncated normal distribution with $\mu = 0$ and $\sigma = \sqrt{2/\text{Indim}}$ where 'Indim' is the number of input units in the weight tensor; the latter draws samples from a uniform distribution within $[-\text{bound}, +\text{bound}]$, where $\text{bound} = \sqrt{(6/(\text{Indim} + \text{Outdim}))}$, with 'Indim' and 'Outdim' indicating the dimensions of the hidden layer and of the following hidden layer, respectively. Additionally, the stochastic gradient descent optimisation algorithm with learning rate 0.0001 is also considered. The number of epochs (with early stopping) analysed are 100 (for the shallow network) and 600 (for the multi-layer NNs). We also consider the following dropout rates (q): 0.01, 0.05, and 0.1.

From Table 1, it is possible to observe that if both methodologies are able to correctly quantify the uncertainty around the predictions of the single and multi-layer NNs across the five pollutants, it is, however, not possible to identify a methodology that consistently outperforms another in terms of out-of-sample predictive accuracy. Additionally, as also shown in Figures 1–7, the two methodologies allow drawing similar conclusions regarding the behaviour of the five pollutants through time.[11]

Table 2 in Levinson and O'Brien (2019) shows that the EECs are constructed considering both numerical and categorical variables (e.g. control variables for age, household size, marital status and indicators for race, education and regional location). To guarantee a proper training of the ReLU neural network, a feature-wise normalisation for the numerical variables—consisting on transforming the observations into zero-mean and unit SD random variables–is performed. For the categorical variables, Levinson and O'Brien (2019) define a separate indicator for each category; being this approach equivalent to the one-hot-encoding procedure (see James et al., 2013), we have applied the same transformation in treating the categorical variables for the correct fitting of the NN. Finally, 85% of the data are used for training the network and the remaining 15% as test set. We focus on the years 1984 and 2012 to assess the evolution of the EECs over time. The optimal combination of structure and hyper-parameters of both single and multi-layer neural networks used for the year 1984 and 2012 are reported in Table 1 with the relative out-of-sample accuracy measures defined by MAE, MSE and the empirical coverage probabilities, Cov_{95}, obtained at a 95% confidence level. Additionally, the Adam optimiser with He normal initialiser was selected as the best optimiser across all pollutants and years considered. The out-of-sample empirical coverage reported in the table is close to the nominal level at which the prediction intervals are constructed, implying the suitability of the intervals out of sample. The results in Table 1 thus convey that the data-based uncovered relationship between the pollution content of US household consumption and income obtained using NNs is statistically robust.

We now turn to examine whether the increasing and concave relationship uncovered by Levinson and O'Brien (2019) is also obtained when estimated nonparametrically using NNs. Figures 1, 2, 3, 4, 5 report the EECs constructed with both MC dropout and EN networks with $d = 2$ and 18 for the five major air pollutants. The panels in each figure are constructed as follows: the observed income is divided into 100 groups of the same size. We compute the mean of the predicted

[11] Given the lower memory (Pomponi et al., 2021) and computational requirements associated with MC dropout, these results would suggest that in case of restrictions in computational power, MC dropout should be preferred over the EN network approach when constructing EEC.

TABLE 1 The optimal neural networks' parameters with relative out-of-sample accuracy measures defined by MSE, MAE and Cov_{95} (empirical coverage probabilities at 95% confidence level)

	Learning rate	Structure	q	T	MAE	MSE	Cov_{95}
PM10							
Year 1984 ($d = 18$)							
MC Dropout	0.003	[78, 36, 36]	0.05	70	2.8594	18.3263	0.95
Extra-network	0.003	[78, 36, 36]	0.05	70	2.9897	19.6456	0.96
Year 1984 ($d = 2$)							
MC Dropout	0.003	[5]	0.05	70	3.6276	27.3797	0.96
Extra-network	0.003	[5]	0.05	70	3.6723	27.5817	0.96
Year 2012 ($d = 18$)							
MC Dropout	0.003	[54, 36]	0.05	70	2.5912	13.2871	0.96
Extra-network	0.003	[54, 36]	0.05	70	2.6437	13.8489	0.96
Year 2012 ($d = 2$)							
MC Dropout	0.003	[5]	0.05	70	2.9997	16.6166	0.96
Extra-network	0.003	[5]	0.05	70	3.0418	17.1902	0.96
CO							
Year 1984 ($d = 18$)							
MC Dropout	0.005	[40, 36]	0.05	70	12.4441	388.4378	0.95
Extra-network	0.005	[40, 36]	0.05	70	12.3746	411.5797	0.95
Year 1984 ($d = 2$)							
MC Dropout	0.005	[5]	0.05	70	14.4730	479.8068	0.94
Extra-network	0.005	[5]	0.05	70	14.4079	494.3869	0.94
Year 2012 ($d = 18$)							
MC Dropout	0.005	[78, 36, 36]	0.05	70	9.0952	236.0679	0.96
Extra-network	0.005	[78, 36, 36]	0.05	70	9.2170	234.8622	0.97
Year 2012 ($d = 2$)							
MC Dropout	0.005	[5]	0.05	70	9.8253	260.4555	0.96
Extra-network	0.005	[5]	0.05	70	9.5545	281.7539	0.96
SO_2							
Year 1984 ($d = 18$)							
MC Dropout	0.005	[54, 36]	0.01	70	30.6778	1844.7867	0.96
Extra-network	0.005	[54, 36]	0.01	70	32.0010	2042.0742	0.95
Year 1984 ($d = 2$)							
MC Dropout	0.005	[5]	0.05	70	37.2184	2810.3498	0.96
Extra-network	0.005	[5]	0.05	70	38.1013	2901.2416	0.95

(Continues)

TABLE 1 (Continued)

	Learning rate	Structure	q	T	MAE	MSE	Cov_{95}
Year 2012 ($d = 18$)							
MC Dropout	0.005	[40, 36]	0.05	70	32.3183	1931.8550	0.96
Extra-network	0.005	[40, 36]	0.05	70	32.5744	1945.6825	0.95
Year 2012 ($d = 2$)							
MC Dropout	0.005	[5]	0.05	70	35.7883	2291.4632	0.97
Extra-network	0.005	[5]	0.05	70	35.8810	2318.8336	0.97
NO_x							
Year 1984 ($d = 18$)							
MC Dropout	0.005	[54, 36]	0.05	70	18.8255	643.9020	0.95
Extra-network	0.005	[54, 36]	0.05	70	19.2223	707.9086	0.94
Year 1984 ($d = 2$)							
MC Dropout	0.003	[5]	0.01	70	23.0578	985.8477	0.96
Extra-network	0.003	[5]	0.01	70	23.1836	1029.5816	0.95
Year 2012 ($d = 18$)							
MC Dropout	0.005	[54, 36]	0.1	70	17.5517	541.5600	0.97
Extra-network	0.005	[54, 36]	0.1	70	17.7210	581.2127	0.96
Year 2012 ($d = 2$)							
MC Dropout	0.005	[5]	0.05	70	19.7508	695.7539	0.96
Extra-network	0.005	[5]	0.05	70	19.9146	699.3635	0.96
VOC							
Year 1984 ($d = 18$)							
MC Dropout	0.003	[78, 36, 36]	0.05	70	5.4403	76.1747	0.96
Extra-network	0.003	[78, 36, 36]	0.05	70	5.2475	76.1247	0.95
Year 1984 ($d = 2$)							
MC Dropout	0.005	[5]	0.05	70	6.5387	96.0701	0.95
Extra-network	0.005	[5]	0.05	70	6.4422	97.8181	0.95
Year 2012 ($d = 18$)							
MC Dropout	0.003	[78, 36, 36]	0.05	70	3.5214	33.7558	0.95
Extra-network	0.003	[78, 36, 36]	0.05	70	3.7183	33.5824	0.96
Year 2012 ($d = 2$)							
MC Dropout	0.005	[5]	0.05	70	4.0099	40.1979	0.96
Extra-network	0.005	[5]	0.05	70	4.0851	41.0851	0.96

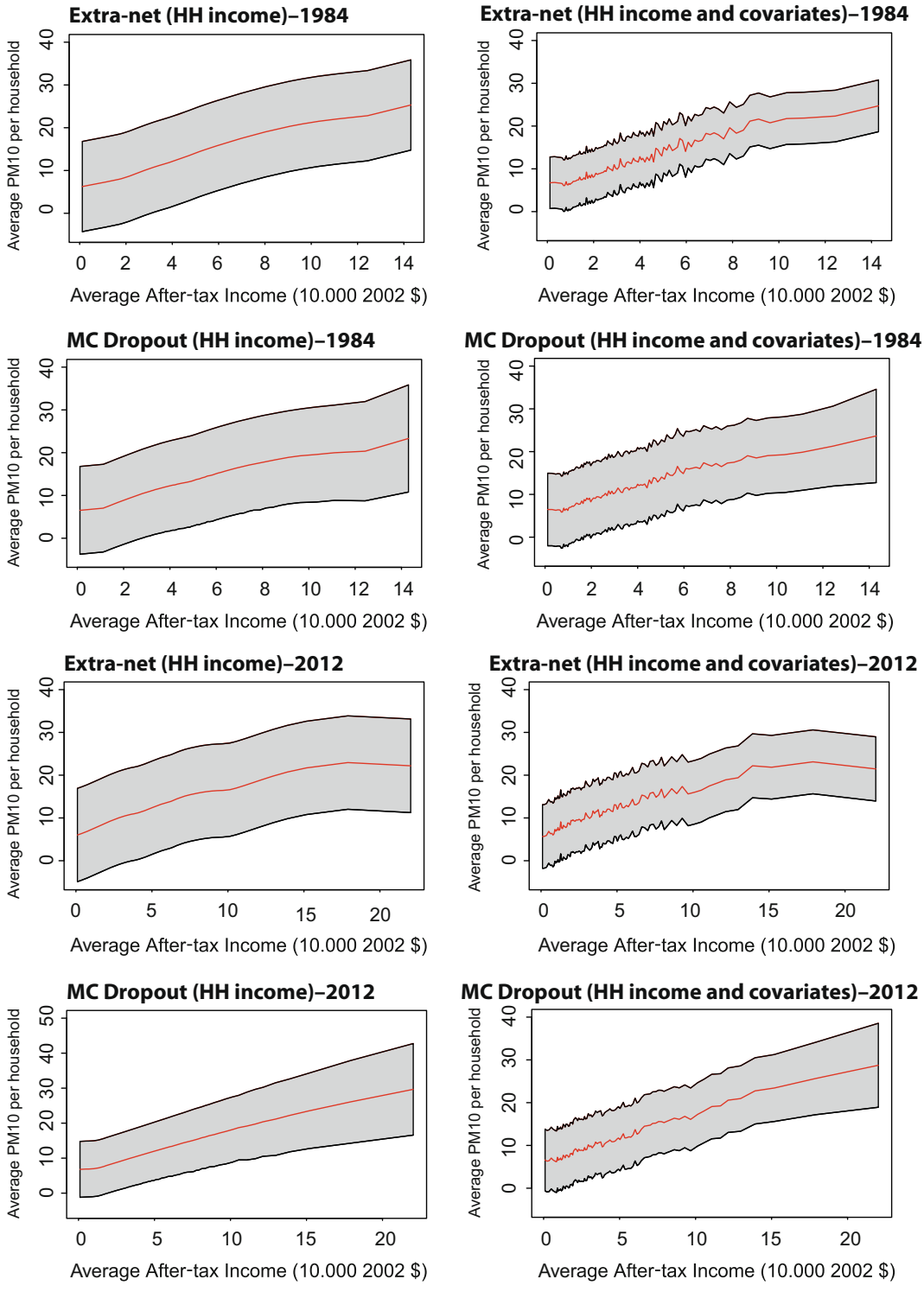

FIGURE 1 Point estimates and 0.95 prediction intervals of particulates smaller than 10 microns

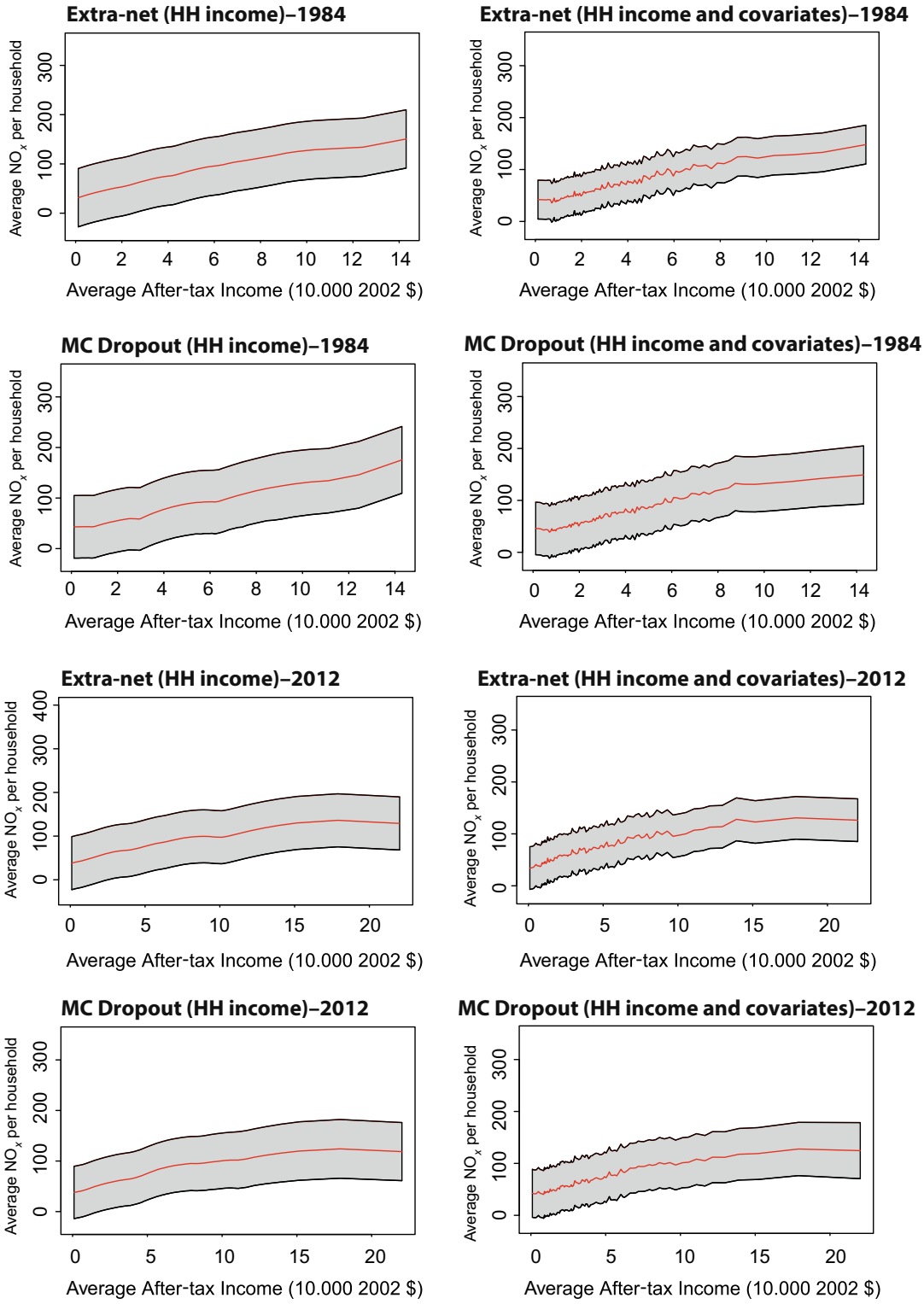

FIGURE 2 Point estimates and 0.95 prediction intervals of nitrogen oxides

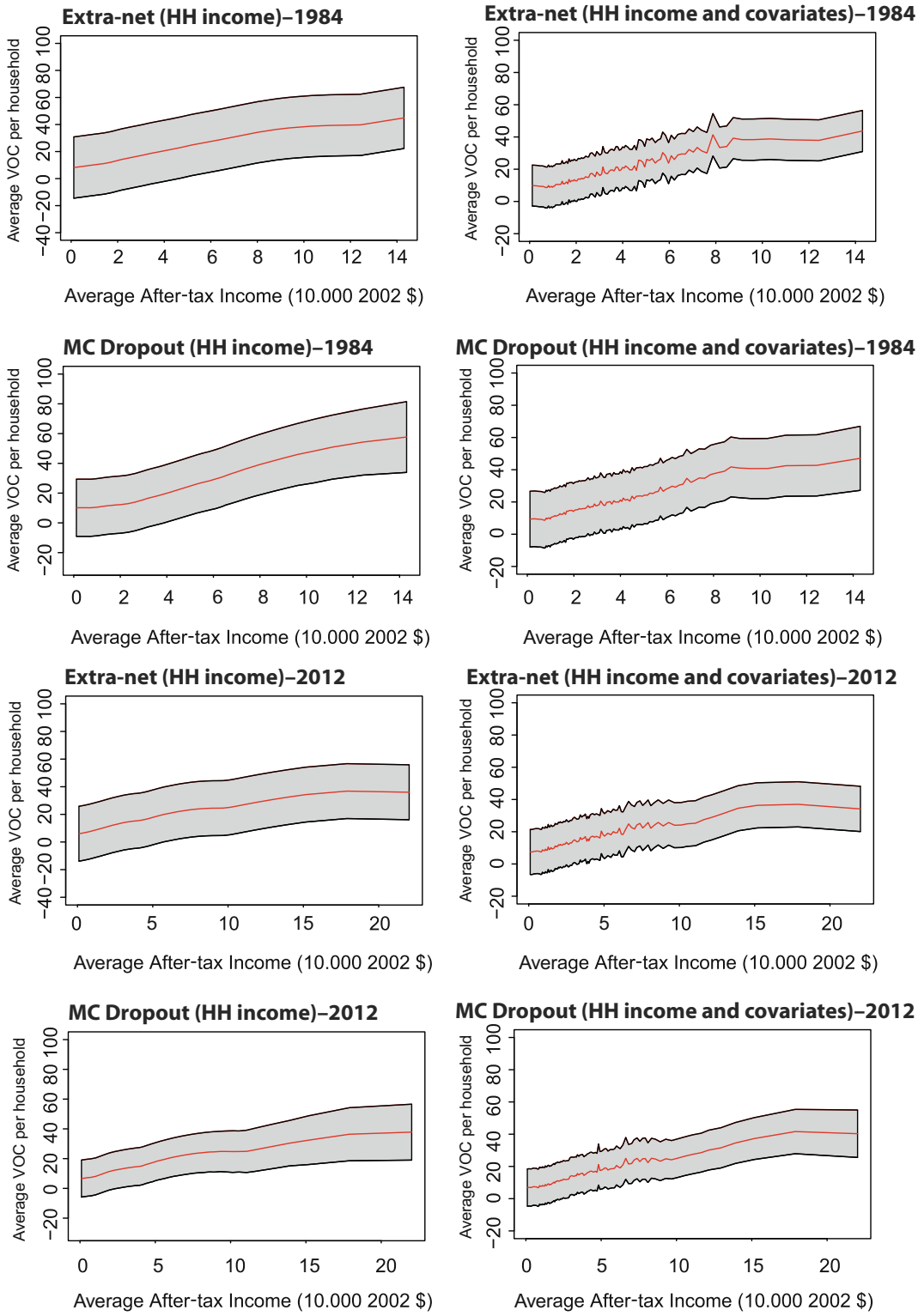

FIGURE 3 Point estimates and 0.95 prediction intervals of volatile organic compounds

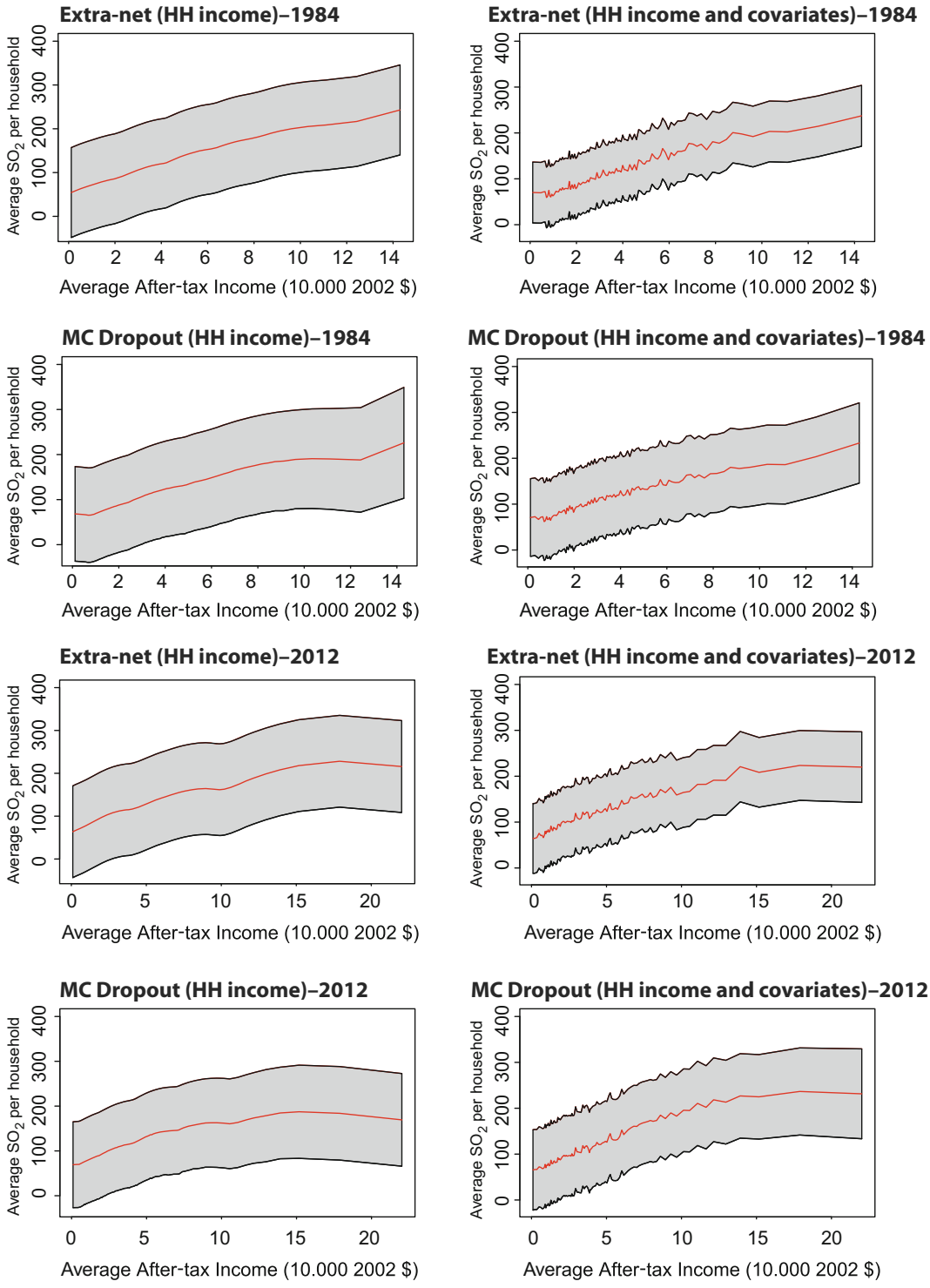

FIGURE 4 Point estimates and 0.95 prediction intervals of sulphur dioxide

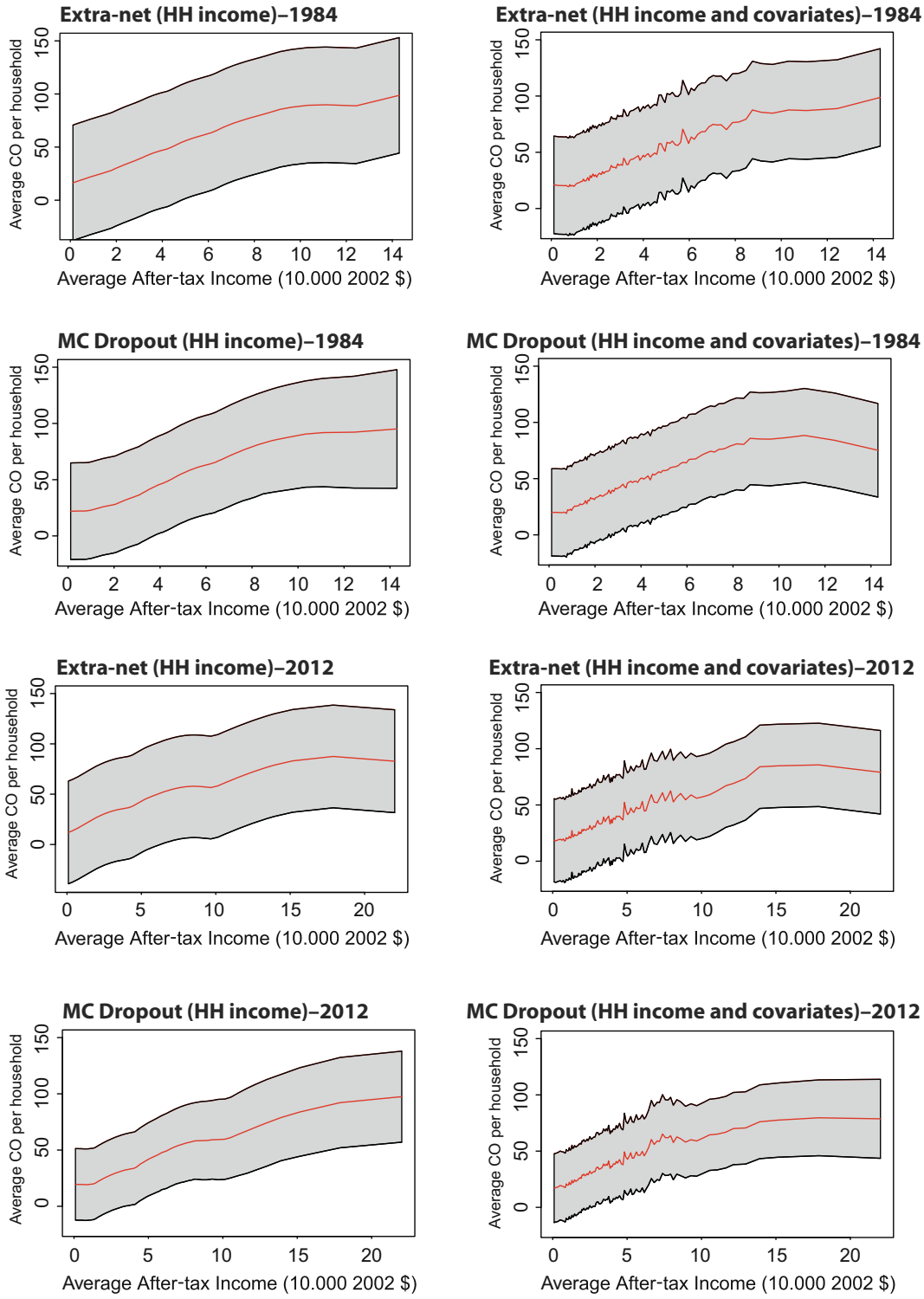

FIGURE 5 Point estimates and 0.95 prediction intervals of carbon monoxide

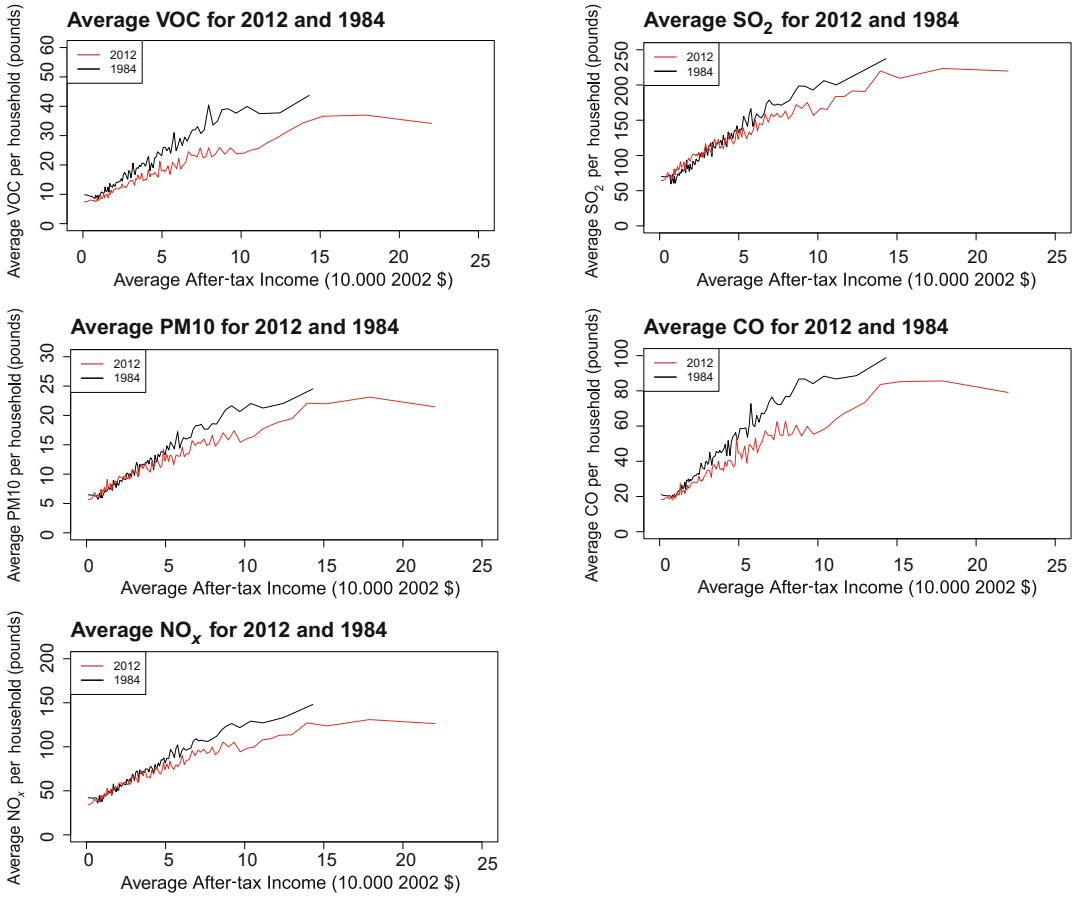

FIGURE 6 Point estimates for the five pollutants obtained from the extra neural network approach

pollution and the upper and lower bounds of the prediction intervals for each group and pollutant. The mean values are then plotted. Our estimates of the EECs constructed with both MC dropout and the EN net approach provide further empirical support to the results reported in figure 3 of Levinson and O'Brien (2019) for the PM10. In particular, the shape of the predicted curves and the associated intervals suggest an increasing and concave relationship between the variables under study. This relationship is uniform across values of household income. More formally, the upper bound of the 95% prediction interval can be used as cut-off value of a statistical test to determine whether the slope of the curve is greater than one and whether its second derivative is negative. The analysis of the prediction intervals obtained from the EN network approach shows that income elasticity is smaller than one across all values of household income. This is particularly the case for the year 2012 and holds across the five pollutants. Although the prediction intervals obtained by MC dropout are not as conclusive as those from the EN nets methodology, overall, there is clear empirical evidence in support of a concave relationship between the pollution content in consumption and household income for the different pollutants considered. These results show that—without imposing any functional form between pollution and household income—richer households pollute more, the pollution content of consumption increases at a lower rate than income, and that the pollution content of consumption grows at a decreasing rate.

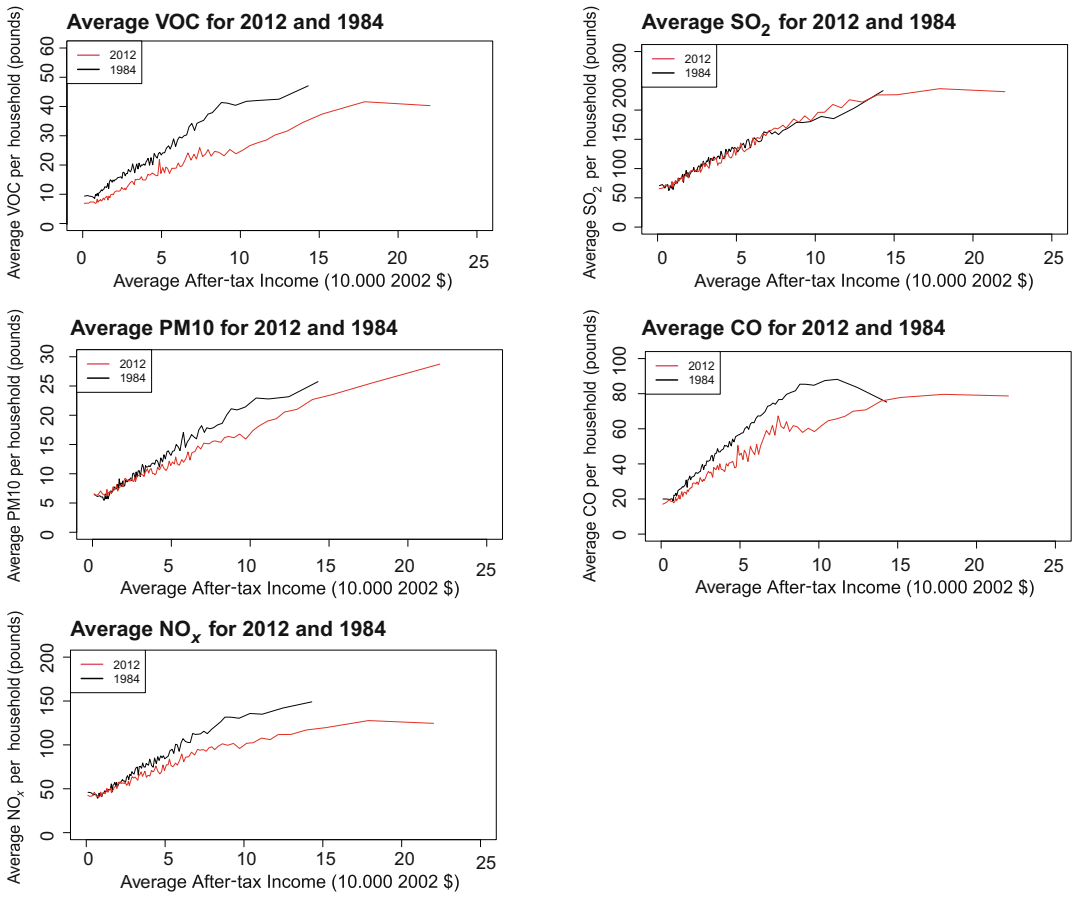

FIGURE 7 Point estimates for the five pollutants obtained from the Monte Carlo Dropout

One salient feature of the estimated EECs obtained under the multi-layer NN corresponds to the year 2012. Compared to 1984, the estimated 2012 EECs display a lower pollution intensity of consumption for US top household earners than for middle-to-high ones. This empirical finding holds for the five pollutants and is more apparent when the multi-layer neural network model and corresponding prediction intervals are obtained using the EN net approach. The effect is weaker if the model is estimated using MC dropout methods but is still apparent in some cases. For some pollutants such as NO_x and CO, we observe a decrease in the level of pollution compared to medium-high income earners suggesting that top earners pollute less than households in the middle to upper range of the household income distribution. Levinson and O'Brien (2019) find some evidence of this phenomenon for 2012 using the simple model with income and income squared but not in the model with multiple covariates. In their model, the quadratic nature of the parametric regression determines to a large extent the overall shape of the relationship between income and pollution, therefore, it is not surprising that the quadratic model captures these effects. There is also the possibility of omitted variable bias in their model although this is ruled out by the authors. In contrast, we find this quadratic effect in the full multi-layer NN with 18s covariates as well as in the simple model. In this nonparametric setting, the model does not impose any structure on the relationship between the variables but, nevertheless, we uncover an inverted U-shape

TABLE 2 The test statistic and the *p*-value of a DM test (on the out-of-sample squared residuals) of a NN model fitted using either Monte Carlo (MC) dropout or the Extra-nets algorithm against the quadratic model considered by Levinson and O'Brien (2019)

Year = 1984					
	PM10	**CO**	**VOC**	**NO_x**	**SO_2**
DM test statistic (EN)	2.8477	2.7741	4.4864	1.9978	2.0995
p-value	(0.0023)	(0.0029)	(<.0001)	(0.0231)	(0.0181)
DM test statistic (MC)	3.1781	2.1401	3.2961	1.4769	2.8909
p-value	(0.0008)	(0.0163)	(0.0005)	(0.0702)	(0.0020)
Year = 2012					
	PM10	**CO**	**VOC**	**NO_x**	**SO_2**
DM test statistic (EN)	3.5784	2.2643	−1.1156	0.7841	2.8625
p-value	(0.0002)	(0.01198)	(0.8674)	(0.2167)	(0.0022)
DM test statistic (MC)	2.0600	0.0122	3.6497	2.4588	3.5089
p-value	(0.01994)	(0.4951)	(0.0001)	(0.0071)	(0.0002)

for the EEC at top income levels that contrasts with standard Engel curves relating consumption and income, holding prices constant.

As Levinson and O'Brien (2019) point out, movements along EECs depend on underlying preferences of richer households relative to poorer households, all else equal. They are independent of any particular environmental policy intervention. In this sense, movements along an EEC reflect a shift in preferences of top earners towards low-polluting goods. A possible explanation of the drop in household pollution for top earners observed in 2012, but not in 1984 and beyond, is due to recent concerns on global warming and climate change. Income is positively correlated with pollution through an increase in household's consumption, however, our NN models uncover a threshold level beyond which the increase in income is not corresponded by an increase in pollution, which suggests that the consumption pattern of top earners is cleaner.[12]

Although non-parametric methods do not allow us to be conclusive about whether this novel result is due to their additional flexibility along a specific dimension (i.e. 'non-linearities' in household income) or between different dimensions (i.e. capturing interactions between household income and additional covariates), we can restrict ourselves to the specific case for which Levinson and O'Brien (2019) find some evidence of this phenomenon in 2012, and compare statistically the predictive performance of our ML models against their quadratic specification. Implementing a Diebold and Mariano (1995) test that evaluates the predictive accuracy of each model using income as single regressor, we test the following hypothesis:

$$\mathcal{H}_0 : \text{MSPE}_{\text{nn}} \geq \text{MSPE}_{\text{quad}}, \tag{23}$$

with MSPE_{nn} denoting the MSPE of NN models and $\text{MSPE}_{\text{quad}}$ the analogous measure for the quadratic econometric specification in Levinson and O'Brien (2019). The rejection of the null

[12]An illustrative example offered by casual evidence would be the use by top earners of cleaner sources of energy consumption for heating and power generation or the use of electric cars. The access to these sources of energy and mobility is more expensive than standard methods based on fossil fuels but result in less-polluting consumption patterns.

hypothesis $\mathcal{H}_0$ implies that the predictions of the NN model are superior in terms of MSPE. The results, reported in Table 2, clearly highlight the outperformance in terms of out-of-sample MSPEs of NN-based predictions over the parametric model proposed by Levinson and O'Brien (2019) in both 1984 and 2012. Absent other covariates, we interpret these results as providing support to the 'additional flexibility' hypothesis along the income dimension that NN methods allow. Yet, we cannot rule out that 'interactions between household income and other covariates' are also responsible.[13]

Finally, Figures 6 and 7 present a comparison of the EECs across years (1984 vs. 2012) for the five pollutants using both NN models. The results provide further support to the empirical insights of Levinson and O'Brien (2019). EECs shift downwards and become more concave over time indicating an overall improvement on the pollution content of households' consumption and also sizeable differences in the intensity of pollution across income levels, which could be explained, at least in part, by shifts in households' preferences towards goods with a lower pollution content.

5 | CONCLUSIONS

Empirical researchers have shown that, despite the economic growth that has characterised the United States in the past 30 years, US households' pollution content of consumption has steadily decreased. Levinson and O'Brien (2019)—by estimating EECs—are able to analyse this relationship. These authors show how the overall pollution in the United States has not increased proportionally with economic growth partially due to changes in the composition of US households' consumption baskets towards less-polluting goods and services.

The present paper further validates the aforementioned empirical findings by adopting NNs to estimate the EECs and associated prediction intervals. The different EECs are constructed using the EN network algorithm recently proposed in Mancini et al. (2021). When only income-related information is considered, a single-layer NN with five hidden nodes is fitted; conversely, when the wider household-specific information set is taken into account, multi-layer NN models are more suitable and provide better fit. Furthermore, recent advances in neural network models allow us to make statistical inference about the pointwise predictions defining the EECs through the construction of prediction intervals. These intervals confirm the empirical findings in Levinson and O'Brien (2019) suggesting that the relationship between after-tax household income and pollution is increasing and concave, that the elasticity of income is lower than one, and that there exists a downward shift in the relationship between the variables when comparing the years 1984 and 2012 .

Importantly, deploying single and multi-layer NNs allows us to uncover an interesting phenomenon for top income earners. For the year 2012, the EEC peaks and then decreases as income reaches the upper range of the household income distribution. This phenomenon is observed for the five pollutants under investigation but is more pronounced for NO_x and CO. In these cases we observe a reduction in the pollution content of consumption compared with medium-to-high income earners, suggesting that top earners pollute less than households right beneath them in the household income distribution. These findings contrast sharply with the strictly increasing

[13]We have also performed an additional Diebold–Mariano test comparing the predictive ability of the NN models with income only and the NN models with 18 covariates. Unsurprisingly, the model with 18 covariates overwhelmingly outperforms the reduced model in terms of predictive ability. Results are not reported for space considerations but are available from the authors upon request.

relationship commonly found for the standard Engel curve between expenditures on normal goods and household income, while providing additional support to the deployment of flexible estimation methods to uncover EECs.

FUNDING INFORMATION

Tullio Mancini acknowledges financial support from the University of Southampton Presidential Scholarship and Jose Olmo from 'Fundación Agencia Aragonesa para la Investigación y el Desarrollo' and project PID2019-104326GB-I00 from the Spanish Secretary of Science and Innovation.

DATA AVAILABILITY STATEMENT

Data for this article is publicly available from a repository accessible from The Review of Economics and Statistics.

ORCID

Jose Olmo https://orcid.org/0000-0002-0437-7812

REFERENCES

Burtraw, D., Sweeney, R. & Walls, M. (2009) The incidence of US climate policy: alternative uses of revenues from a cap-and-trade auction. *National Tax Journal*, 62(3), 497–518.

Calvo-Pardo, H. F., Mancini, T. & Olmo, J. (2021) *Optimal deep neural networks by maximization of the approximation power*. Available at: https://ssrn.com/abstract=3578850 or https://doi.org/10.2139/ssrn.3578850 [Accessed 15th October 2021].

Copeland, B. & Taylor, M.S. (2005) *Trade and the environment: theory and evidence*. Princeton: Princeton University Press.

Cortes-Ciriano, I. & Bender, A. (2019) Reliable prediction errors for deep neural networks using test-time dropout. *Journal of Chemical Information and Modeling*, 59(7), 3330–3339.

Cybenko, G. (1989) Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4), 303–314.

Denker, J.S. & LeCun, Y. (1991) Transforming neural-net output levels to probability distributions. In: Lippmann, R., Moody, J. & Touretzky, D. (Eds.) *Advances in neural information processing systems*. Denver: Morgan Kaufmann, pp. 853–859.

Devieaux, R.D., Schumi, J., Schweinsberg, J. & Ungar, L.H. (1998) Prediction intervals for neural networks via nonlinear regression. *Technometrics*, 40(4), 273–282.

Diebold, F. & Mariano, R. (1995) Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13(3), 253–263.

Dipu Kabir, H.D., Khosravi, A., Hosen, M.A. & Nahavandi, S. (2018) Neural network-based uncertainty quantification: a survey of methodologies and applications. *IEEE Access*, 6, 36218–36234.

El Karoui, N. & Purdom, E. (2018) Can we trust the bootstrap in high-dimensions? the case of linear models. *The Journal of Machine Learning Research*, 19(1), 170–235.

Engel, E. (1895) Das Lebenskosten belgischer arbeiterfamilien fruher und jetzt. *Bulletin de Institut International de Statistique*, 9, 1–124.

Gal, Y. & Ghahramani, Z. (2016) Dropout as a bayesian approximation: Representing model uncertainty in deep learning. *International conference on machine learning*, New York, pp. 1050–1059.

Geurts, P., Ernst, D. & Wehenkel, L. (2006) Extremely randomized trees. *Machine Learning*, 63(1), 3–42.

Goodfellow, I., Bengio, Y. & Courville, A. (2016) *Deep learning*. Cambridge/London: MIT Press.

Grainger, C.A. & Kolstad, C.D. (2010) Who pays a price on carbon? *Environmental and Resource Economics*, 46, 359–376.

Grossman, G. & Krueger, A. (1995) Economic growth and the environment. *Quarterly Journal of Economics*, 110, 353–377.

Hassett, K., Mathur, A. & Metcalf, G. (2009) The incidence of a U.S. carbon tax: a lifetime and regional analysis. *Energy Journal*, 30, 155–177.

Hayashi, N. (2020) Variational approximation error in non-negative matrix factorization. *Neural Networks*, 126, 65–75.

Heskes, T. (1997) Practical confidence and prediction intervals. In: Mozer, M.C. (Ed.) *Advances in neural information processing systems*. Cambridge: MIT Press, pp. 176–182.

Hornik, K. (1991) Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2), 251–257.

Hwang, J.G. & Ding, A.A. (1997) Prediction intervals for artificial neural networks. *Journal of the American Statistical Association*, 92(438), 748–757.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013) *An introduction to statistical learning: with application in R*. New York: Springer.

Kendall, A. & Gal, Y. (2017) What uncertainties do we need in Bayesian deep learning for computer vision? *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 5580–5590.

Lakshminarayanan, B., Pritzel, A. & Blundell, C. (2017) Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, Long Beach, pp. 6402–6413.

Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D. and Batra, D. (2015) Why M heads are better than one: training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314*.

Levinson, A. & O'Brien, J. (2019) Environmental Engel curves: indirect emissions of common air pollutants. *Review of Economics and Statistics*, 101(1), 121–133.

Lu, Z., Pu, H., Wang, F., Hu, Z. & Wang, L. (2017) The expressive power of neural networks: a view from the width *Advances in neural information processing systems*, Long Beach, pp. 6231–6239.

Mancini, T., Calvo-Pardo, H.-F. & Olmo, J. (2021) Extremely randomized neural networks for constructing prediction intervals. *Neural Networks*, 144, 113–128.

Metcalf, G. (1999) A distributional analysis of green tax reforms. *National Tax Journal*, 52, 655–682.

Montufar, G.F., Pascanu, R., Cho, K. & Bengio, Y. (2014) On the number of linear regions of deep neural networks. *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2924–2932.

Pomponi, J., Scardapane, S. & Uncini, A. (2021) Structured ensembles: an approach to reduce the memory footprint of ensemble methods. *Neural Networks*, 144, 407–418.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. (2014) Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1), 1929–1958.

Stone, C.J. (1980) Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6), 1348–1360.

Tibshirani, R. (1996) A comparison of some error estimates for neural network models. *Neural Computation*, 8(1), 152–163.

Zhou, Z.H. (2012) *Ensemble methods: foundations and algorithms*. New York: Chapman and Gall/CRC Press.

How to cite this article: Mancini, T., Calvo-Pardo, H. & Olmo, J. (2022) Environmental Engel curves: A neural network approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 1–26. Available from: https://doi.org/10.1111/rssc.12588