



Are Climate Model Forecasts Useful for Policy Making? Effect of Variable Choice on Reliability and Predictive Validity

Kesten C. Green¹

Willie Soon²

¹UniSA Business and Ehrenberg-Bass Institute, University of South Australia
Commerce CWE-31, GPO Box 2471, Adelaide SA 5001, Australia

²Institute of Earth Physics and Space Science (HUN-REN EPSS), H-9400 Sopron, Hungary
Center for Environmental Research and Earth Sciences, Salem, Massachusetts, 01970, USA.

SCC Publishing
Michelets vei 8 B
1366 Lysaker Norway

ISSN: 2703-9072

Correspondence:
kesten.green@unisa.edu.au

Vol. 5.1 (2025)

pp. 59 - 85

Abstract

For a model to be useful for policy decisions, statistical fit is insufficient. Evidence that the model provides out-of-estimation-sample forecasts that are more accurate and reliable than those from plausible alternative models, including a simple benchmark, is necessary.

The UN's IPCC advises governments with forecasts of global average temperature drawn from models based on hypotheses of causality. Specifically, manmade warming principally from carbon dioxide emissions (Anthro) tempered by the effects of volcanic eruptions (Volcanic) and by variations in the Sun's energy (Solar). *Out-of-sample* forecasts from that model, with and without the IPCC's favoured measure of Solar, were compared with forecasts from models that excluded human influence and included Volcanic and one of two independent measures of Solar. The models were used to forecast Northern Hemisphere land temperatures and—to avoid urban heat island effects—rural only temperatures. Benchmark forecasts were obtained by extrapolating estimation sample median temperatures.

The independent solar models reduced forecast errors relative to those of the benchmark model for all eight *combinations* of four estimation periods and the two temperature variables tested. The models that included the IPCC's Anthro variable reduced errors for only three of the eight combinations and produced extreme forecast errors from most model estimation periods. The mean correlation between estimation sample statistical fit and forecast accuracy was -0.30.

Further tests might identify better models: Only one extrapolation model and only two of many possible independent solar models were tested, and combinations of forecasts from different methods were not examined.

The anthropogenic models' unreliability would appear to void policy relevance. In practice, even the models validated in this study may fail to improve accuracy relative to naïve forecasts due to uncertainty over the future causal variable values. Our findings emphasise that out-of-sample forecast errors, *not* statistical fit, should be used to choose between models (hypotheses).

Keywords: alternative hypotheses; causal models; climate policy; goodness-of-fit; R-squared; scepticism; scientific method; simplicity (Occam's razor)

Submitted 2025-03-11, Accepted 2025-05-17, <https://doi.org/10.53234/scc202501/07>

1. Introduction

What has caused changes in annual average temperatures on Earth over recent decades and what, therefore, can we expect in the way of temperature changes over coming decades?

Those are questions that were originally posed by the United Nations Intergovernmental Panel on Climate Change (IPCC) in the Second Assessment Report published in 1996. The first of the questions has been referred to as the “attribution problem” among contributors to the IPCC project (see p. 413 Santer et al., 1996). Proposed answers can be expressed as hypotheses, or mathematical models, of the putative causal relationships.

The second question is a forecasting, or predictive validity, problem that can be answered using a key element of the scientific method. Namely, the testing of multiple reasonable hypotheses of causality (models) against data that were not used in the development of the hypotheses and estimation of the model parameters by forecasting the data that were unknown to the model. Doing so enables the researcher to identify the model with the greatest predictive validity and, potentially, to use that model for making useful forecasts.

One plausible causal variable is the energy flux and flow that the Earth receives from the Sun, which can in turn be thought of as being, broadly, a function of variations in (a) the quantum of solar energy that reaches the Earth’s atmosphere, and the proportions of that energy that (b) reaches the Earth’s surface, and (c) escapes from the Earth’s surface and atmosphere. The energy that reaches and escapes from the Earth’s surface is affected by the composition of the Earth’s atmosphere—gases and aerosols—the composition of which may change.

The IPCC were tasked with identifying the effect of human activity on global and regional surface temperatures. Chapter 8 of the IPCC’s Second Assessment Report—titled, “Detection of Climate Change and Attribution of Causes”—opens with the statement “Since the 1990 IPCC Scientific Assessment considerable progress has been made in attempts to identify an anthropogenic effect on climate” (p. 411, Santer et al., 1996). Other statements in the chapter summary indicate the authors’ reservations regarding the state of knowledge on human influences on climate: e.g., “...large uncertainties still apply to current estimates of the magnitude and patterns of natural climate variability...” and “Our ability to quantify the magnitude of this effect is presently limited by uncertainties in key factors, such as the magnitude of longer-term natural variability and the time-evolving patterns of forcing and response to changes in greenhouse gases, aerosols and other human factors” (p. 411, Santer et al., 1996).

In their attempts to achieve the IPCC objective of identifying a human cause for temperature changes—specifically “global warming”—the IPCC researchers have framed the problem as one of “attributing” changes in the Earth’s temperature to the respective contributions of putative anthropogenic (“Anthro”)—principally carbon dioxide emissions altering the composition of the atmosphere—and natural influences—principally aerosols from volcanic eruptions altering the composition of the atmosphere (“Volcanic”), and total solar irradiance, or TSI, variations (“Solar”). Given the task they were set, the IPCC researchers have devoted much of their efforts into developing estimates of the Anthro variable.

The IPCC’s most recent, AR6, report (IPCC, 2021) only considered one estimate of Solar for the purpose of attribution (Matthes et al., 2017) and made no allowance for the effect of urban heat islands on the temperature measures they used (Connolly et al., 2021, 2023; Soon et al., 2023). Moreover, a study of the statistical attribution or “fingerprinting” approach used by IPCC researchers (e.g., Allen and Tett, 1999; Hasselmann, et al., 1995; Hegerl et al., 1997; Santer et al., 1995) concluded that the approach was invalid (McKittrick, 2022). The IPCC authors’ analyses failed to meet the assumptions of the method they used, and they failed to correctly implement the method.

In sum, the objective given to the IPCC researchers and the approach that they have taken suggests that plausible alternative hypotheses on the causes of terrestrial temperature changes may not have been adequately tested, as is required by the scientific method (Armstrong and Green, 2022). That concern is consistent with Armstrong and Green’s (2022) observation that government sponsorship of research can create incentives that may influence researchers’ choices of hypotheses to test and how they test them.

1.1 Alternative hypotheses on Solar

To address the first of the foregoing limitations in the IPCC attribution studies—failure to fairly test alternative TSI estimates—Connolly et al. (2021, 2023) comprehensively reviewed alternative estimates of TSI covering the 169 years from 1850 to 2018. In addition to the Matthes, et al. (2017) TSI estimates series used by the IPCC (2021)—henceforth “IPCC Solar”—Connolly et al. (2023) identified 27 alternative Solar time series.¹

The alternative estimates of Solar correlate quite well with the TSI used in the AR6 report—Pearson’s r values range between 0.39 and 0.97 with a median of 0.82—but the degree of TSI variation in Watts per square metre (Wm^{-2}) differs considerably between the estimates. The ranges of the individual alternative TSI estimate series vary between 0.49 and 4.64 Wm^{-2} , with a median range of 1.77 Wm^{-2} , while IPCC Solar has a range of only 0.19 Wm^{-2} .

In this study, we consider two plausible TSI reconstructions from Connolly et al. (2023). Those from Hoyt and Schatten (1993) and from Bard et al. (2000), which Connolly et al. (2023) updated to the year 2018². The former TSI record (“H1993 Solar”) was based on the so-called multi-proxy—i.e., equatorial solar rotation rate, sunspot structure, the decay rate of individual sunspots, the number of sunspots without umbrae, and the length and decay rate of the 11-yr sunspot activity cycle—reconstruction of the solar irradiance history.

H1993 Solar was previously considered and endorsed by the IPCC’s AR4 report (IPCC, 2007; see “Supplementary Materials for Chapter 9” of Working Group One report). In contrast to IPCC Solar, the range of values taken on by H1993 Solar is the second largest of the 27 alternative TSI estimates at 4.2 Wm^{-2} . Its correlation with IPCC Solar was the 23rd largest, or fourth smallest, at 0.62.

The latter TSI reconstruction (“B2000 Solar”) was based on cosmogenic isotope measurements of ^{14}C in tree rings and ^{10}Be in polar ice cores. The range of values taken on by B2000 Solar is the 11th largest of the 27 alternative TSI estimates at 2.1. Its correlation with IPCC Solar was the lowest of the 27 alternatives at 0.39. The correlation between B2000 Solar and the AR4-endorsed H1993 Solar, on the other hand, was 0.77, which is the fourth highest correlation among the correlations of the TSI alternatives with H1993.

Note that the alternative TSI time series—H1993 Solar and B2000 Solar—are both independent of each other and from the temperature data used in the analysis presented in this paper. The two records are derived from physically distinct measures: the former from solar activity measures (mainly sunspots), and the latter from incoming cosmic ray data. Temperature data are from terrestrial weather station records.

1.2 Alternative hypotheses on temperature estimation

The IPCC’s attribution studies do not account for the *direct* effects of human activities on local temperatures (heat islands)—the second weakness addressed in this study. For example, heating and cooling of building interiors, electricity generation, manufacturing, freight and transport, asphalt and concrete, and where vegetation and open water have been removed or added. Where temperature readings are taken close to such human sources of heat or absence of natural cooling,

¹ Connolly, *et al.*’s (2023) data set is available at <https://zenodo.org/records/8225275> (DOI 10.5281/zenodo.8225274), and is the only data used in the study presented in this paper.

² Connolly, *et al.* (2023) used version 3 of the Global Historical Climatology Network (GHCN) data, which ends in 2018.

they cannot properly reflect the individual effects of human emissions of carbon dioxide, etc., that the IPCC are concerned about (their Anthro variable), the Volcanic variable, and TSI.

To address this second limitation in the IPCC attribution studies, Connolly et al. (2021, 2023) developed four alternative estimates of surface temperatures that were intended to avoid heat island effects. They were based on rural only weather station readings, sea surface temperature readings, tree-ring width measurements, and glacier length measurements. For comparison with the approach used by the IPCC, they also developed an all-land temperature estimates series for the Northern Hemisphere.

Connolly et al.'s (2021, 2023) temperature series were confined to estimates for the Northern Hemisphere due to the relative scarcity of long records from geographically dispersed locations in the Southern Hemisphere. This study uses Connolly et al.'s (2021, 2023) all-land and the alternative rural-land series—"NH All Land Annual Average Temperature" and "NH Rural Land Annual Average Temperature," respectively—in order to test hypotheses on causality. The two series were derived from individual thermometer and weather station readings.³

1.3 Hypothesis validity criterion

In assessing alternative hypotheses on TSI and temperature estimation, Connolly et al. (2021, 2023) were primarily concerned with determining the *plausibility* of alternative models of surface temperature relative to the IPCC hypothesis that temperature changes over recent decades have been in a large part caused by human emissions of carbon dioxide, etc. (Anthro). To do that, they compared the statistical fits (coefficients of determination, or R^2 s) of their alternative models using a diversity of estimates of TSI and of temperature. They found that many of the models that included an alternative Solar estimate, but not the IPCC's Anthro, were as plausible if not more so than the IPCC's Anthro, Volcanic, IPCC Solar formulation when assessed against the R^2 criterion.

Measures of statistical fit such as R^2 do not, however, provide useful information about predictive validity. There are theoretical reasons and empirical evidence in support of that assertion.

For the former—theoretical reasons—consider that correlation is not evidence of causation, and that especially with the plethora of data on many and various phenomena readily available, it is not difficult to achieve a statistical fit measurement that is consistent with discipline norms without consideration of evidence on causality. Take, for example, an ordinary least squares (OLS) regression model of NH All Land Annual Average Temperature as a function of U.S. Postal rates for stamped cards estimated over the period 1873 to 2018—the period of data overlap (Historian, 2023). The adjusted- R^2 of the model is 0.35, which is greater than adjusted- R^2 s of 12 of the 35 proposed causal model formulations listed in Connolly et al.'s Table 1 (2023).

For the latter—empirical evidence on the relationship between statistical fit and validity—see, e.g., Armstrong (2001) on evaluating forecasting methods. See also Soon, Connolly and Connolly's (2015) discussion on four types of correlations: causal, commensal, coincidental and constructional. The IPCC reports' authors are not consistently clear in distinguishing between the four possibilities in the conduct of their "attribution" studies. Given that the collective of IPCC researchers is responsible for constructing the data series that are used in their models, there is a

³ No indirect temperature proxies such as those derived from ice or sea sedimentary cores or from tree-rings were involved in this study: All the temperature records were from instrumental thermometers. We were only concerned with contrasting the outcomes based on the rural-only temperature data with those derived from rural and urban (i.e., "All") temperature data. Another important piece of evidence to consider is that the rural-only temperature record we adopted in this paper compares well with sea surface temperature and two other indirect temperature proxies based on tree-ring and glacier studies. The relationships are documented in Connolly et al. (2023).

possibility that the correlations are to some extent unintentionally constructional, rather than causal.

Comparing R^2 s cannot, therefore, help to answer the second question: “what can we expect in the way of temperature changes over coming decades?”. To answer the second question, we need to know which hypothesis, or model, provides forecasts about data unknown to the model that are more accurate than those from a simple and plausible benchmark model and more accurate than those from alternative plausible models.

This study provides a partial answer to that question by testing some of the model/variable combinations developed by Connolly et al. (2021, 2023) for their predictive validity relative to a simple benchmark model. In accord with good scientific practice, the model/variable combinations were chosen *a priori*, for the reasons described in section 2.1 below, and no other combinations were tested.

1.4 Benchmark hypothesis

To obtain useful scientific findings, a plausible benchmark hypothesis—ideally the likely strongest competitor—is needed. In this study, we needed a model that was consistent with evidence-based forecasting principles. An overarching principle of forecasting—known as The Golden Rule of Forecasting—requires that models should be conservative, by staying close to what is known about the situation of interest and to what is known about forecasting methods (Armstrong, Green, and Graefe 2015). Expressed in the negative, forecasts that the future will be different from what is known about the past and forecasts from methods other than those that have been extensively tested against evidence-based methods and validated using out-of-sample forecasts should be rejected by decision makers.

There is much uncertainty surrounding the causal relationships behind changes in the Earth’s climate in general and temperatures in particular. If that were not the case, there would be no need for “attribution” studies, no need for debates about data, and unexpectedly large or biased deviations from model predictions would not occur. As discussed below, the IPCC anthropogenic variable is a complex compound of 11 putative human influences on global temperatures. Moreover, complex computer climate models have failed to emulate the internal modes of changes and variability of known phenomena such as El-Nino Southern Oscillation (ENSO), Pacific Decadal Oscillation (PDO), and Atlantic Multidecadal Oscillation as well as nearly all the physical-chemical-biological processes on land and in the oceans (see, e.g., Soon et al. 2001; Notz 2015; Eyring et al. 2019; Beverley et al. 2024; Wang et al. 2024).

Particularly with such a complex and uncertain situation, the scientific principle of simplicity—the application of Occam’s razor—is appropriate. A meta-analysis of comparisons between simple and complex forecasting methods found that simple models—often naïvely simple—provided forecasts that were as or more accurate than those from the more complex models that researchers were proposing (Green and Armstrong 2015).

In the light of those forecasting and general scientific principles, it is reasonable to ask: is it possible to forecast climate change better than some historical average or average trend? That question was addressed by Green, Armstrong, and Soon (2009). That paper rejected extrapolating a trend on the basis that apparent trends in historical temperature readings and in various proxy temperature reconstructions reversed directions on all time scales without clear cause.⁴ In other

⁴ Over the past two centuries or so, scientists have warned about the future consequences of recent temperature trends. For example, at the first Earth Day in 1970, Professor Kenneth Watt gave a speech declaring, “The world has been chilling sharply for about twenty years... If present trends continue, the world will be about four degrees colder for the global mean temperature in 1990, but eleven degrees colder in the year 2000. This is about twice what it would take to put us into an ice age” (Bailey, 2000).

words, an increase was more-or-less as likely to be followed by a decrease as by another increase, and vice versa.

The authors instead hypothesised that a simple, even naïve, model that forecasts that future years' temperatures would be the same as the previous year's temperature would be hard to beat to any policy-relevant extent. Their study found that to be the case with, for example, the cumulative absolute error from projecting the IPCC's then 3 °C-per-century "business as usual" warming rate (p. 17, IPCC 1992) out to 2008 from successively updated starting years from 1851 to 1975 to be seven times greater than the error from projecting the previous year's temperature out to 2008 (Green, Armstrong, and Soon 2009).

This study used a variation on the Green, Armstrong, and Soon (2009) naïve no-change (no trend) model. The benchmark model used here forecasts that temperatures will be equal to the median of temperatures in the estimation sample. Medians are a more conservative (see Armstrong, Green, and Graefe, 2015) form of averaging than means, which can be greatly affected by outliers, as with averaging incomes or real estate prices.

1.5 Hypotheses tested

The foregoing discussion suggests the following hypotheses, which are tested in this study.

- H1. Forecasts from causal models will [will not] be usefully more accurate than forecasts from a naïve no-change model.
- H2. Models using variable measures developed independently of the IPCC dangerous manmade global warming hypothesis will [will not] have greater predictive validity.
- H3. The statistical fit of the models (adjusted- R^2) will not [will] be substantively positively related to their predictive validity.
- H4. Models using variable measures developed independently of the IPCC dangerous manmade global warming hypothesis will [will not] be more reliable.⁵

2. Methods

2.1 Models and variables

Ordinary least squares regression models were estimated using data for the eight combinations of forecast variable and causal variables listed in Table 1.

The variables Anthro, Volcanic, and Solar IPCC were nominated in the AR6 report (IPCC, 2021) as the primary causes ("drivers") of changes in the Earth's surface temperature. The variables are measured in terms of "radiative forcings" in Wm^{-2} .

Anthro is the IPCC AR6's composite of 11 proposed anthropogenic influences, which is dominated by atmospheric carbon dioxide and somewhat ameliorated by both the direct—through modulation of the radiation flows—and indirect—through modulation of cloud variables—effects of aerosols. On the other hand, the AR6 report identifies only two "natural forcings": one estimate of the effect of volcanic eruptions (Volcanic), and one estimate of the effect of TSI as described in Matthes, et al. (2017) (here "IPCC Solar").

⁵ Out-of-sample forecast errors of reliable models will exhibit orderly progressions as additional observations are added to the estimation sample, suggesting more realistic representations of causal relationships.

Table 1: Models and Variables

Model Name	Causal variables (Wm ⁻²)			Forecast variable
AVL	Anthro [†]	Volcanic [‡]	-	
AVSL	Anthro [†]	Volcanic [‡]	Solar IPCC [§]	NH All Land Annual Average Temperature Anomaly
S_BVL	Solar B2000 ^{§§}	Volcanic [‡]	-	
S_HVL	Solar H1993 [*]	Volcanic [‡]	-	
AVR	Anthro [†]	Volcanic [‡]	-	
AVSR	Anthro [†]	Volcanic [‡]	Solar IPCC [§]	NH Rural Land Annual Average Temperature Anomaly
S_BVR	Solar B2000 ^{§§}	Volcanic [‡]	-	
S_HVR	Solar H1993 [*]	Volcanic [‡]	-	

[†] *Anthro is a composite of 11 proposed anthropogenic influences dominated by atmospheric carbon dioxide (IPCC 2021).*

[‡] *Volcanic is the estimated effect of volcanic eruptions (IPCC 2021).*

[§] *IPCC Solar is the estimate of TSI preferred in the IPCC (2021) report (Matthes, et al. 2017).*

^{§§} *Solar B2000 is Bard et al. (2000) as updated to 2018 by Connolly et al. (2023).*

^{*} *Solar H1993 is Hoyt and Schatten (1993) as updated to 2018 by Connolly et al. (2023).*

The two alternative measures of Solar identified by Connolly et al. (2023) used in the models tested in this study; namely Solar B2000 and Solar H1993 (Bard, Raisbeck, Yiou, and Jouzel, 2000; Hoyt and Schatten, 1993) were chosen to contrast with the low-variability IPCC Solar estimate of Matthes, et al. (2017) suggested by the CMIP6 historical model runs that were adopted in the AR6 Working Group One report (IPCC 2021).

Finally, the two temperature variables forecast in this study were chosen from the five examined by Connolly et al. (2023) as being the most directly relevant to human experience. The series were estimates of *all* Northern Hemisphere land surface temperatures and estimates of *rural* land Northern Hemisphere temperatures. Both series were based on weather station readings. Two of the three temperature series from Connolly et al. (2023) not considered in this study were proxy measures of temperature based on tree ring data in one case and glacier length in the other. The third series not considered in this study was based on ocean surface temperature readings.

The NH All Land temperature series is representative of the IPCC approach to estimating regional and global temperatures, whereas the NH Rural Land series is an independent attempt to estimate temperature anomalies that are uncontaminated by urban heat island effects (Connolly et al. 2021, 2023; Soon et al. 2023). The distinction is important because the IPCC model of climate change is based on the hypothesis that emissions of infrared active “greenhouse” gasses—mainly carbon dioxide—from human activity causes substantively harmful *global* warming. Urban heat islands are local and are caused by land surface changes and other environmental factors within cities, towns, and suburbs—including heat generated by motors and engines of all kinds—and not from global changes in the concentration of greenhouse gasses. Consequently, out-of-sample forecasts of the NH Rural Land temperature series can be expected to provide more realistic testing of the IPCC, and of the independent, hypotheses about causality. Given that none of the models include an “urban heat island” causal variable, one would expect the more realistic causal models to provide smaller errors in forecasting the Rural series than in forecasting the All Land series.

The data series described above and used in this study are available from Connolly et al. (2023). We followed Connolly et al. (2023) in estimating causal models of temperature anomalies in degrees Celsius against the putative causal variables in Wm⁻² using that data. Appendix A charts the relationships between the alternative temperature measures and the alternative causal variables. Several things stand out. First, the IPCC’s preferred Solar (TSI) variable does not vary; at least not in comparison to the other causal variables considered; in fact, the correlation with

temperature is negative from 1970 (Figure A1 charts C and D). Second, the IPCC’s Anthro variable varied little over the 120 years up to 1970, but after that had a strong positive correlation with temperature implying a dramatically different effect size (Figure A1 charts A and B). Third, the independent Solar variables B2000 and H1993 do vary and do exhibit positive correlations with temperature both pre and post 1970 as one would expect of valid causal variables (Figure A1 charts E through F). Finally, while the NH Rural Land temperatures (Figure A1 charts B, D, F, and H) vary over a greater range than the NH All Land temperatures (Figure A1 charts A, C, E, and G), the patterns of the relationships are similar.

While noting that the time series concept of stationarity is not relevant to the causal model approach that we take in this paper, in deference to a reviewer we provide tests of stationarity and tests of models estimated using stationary data in Appendix B. Although a test of the first 50 years of the available data found the causal variables other than Volcanic failed stationarity tests, estimating causal models from data adjusted to achieve stationarity failed to provide the substantial reductions in out-of-sample forecast errors that the models estimated using the original data presented in the body of this paper achieved.

2.2 Test 1: Predictive validity of causal models and relationship to R^2 [H1, H2, H3]

To obtain out-of-sample forecasts with which to test the predictive validity of the models, the models’ parameters were estimated using OLS regression applied to subsamples of the available historical data from 1850. Three subsamples of 50, 100, and 150 years of data were used, in addition to a subsample of 120 observations. The latter subsample, ending in 1969, was used in the testing because it includes the data leading up to the current manmade global warming alarm. The subsamples can be thought of as representing the data available to forecasters at intervals through history assuming the IPCC’s and others’ estimates had been compiled each year since 1850.

The models were used to forecast the out-of-sample years to 2018, which provided 119, 69, 19, and 49 forecasts, respectively, from each model. The out-of-sample causal variables were *not* forecasted, rather known—measured or estimated—values were used to derive the forecasts.

Accuracy comparisons were made using four measures of out-of-sample forecast errors, the most basic being the median absolute error (*MdAE*) and the interquartile range (*IQR*) of the signed errors, both in °C units.

The other two error measures used were relative measures: the cumulative relative absolute error (*CumRAE*)—also known as the relative mean absolute error (*RelMAE*)—and the unscaled mean bounded relative absolute error or *UMBRAE* (Armstrong and Collopy 1992; Chen, Twycross and Garibaldi 2017). The *CumRAE* is easier to understand, but the *UMBRAE* is preferred as it is based on individual forecast comparisons and so better reflects the typical relative error. The formulas for both are shown below.

$$CumRAE \text{ or } RelMAE = \frac{\sum_{i=1}^n |e_i|}{\sum_{i=1}^n |e_i^*|} \quad (1)$$

$$UMBRAE = \frac{MBRAE}{1 - MBRAE} \quad (2)$$

$$\text{where } MBRAE = \frac{1}{n} \sum_{i=1}^n \frac{|e_i|}{|e_i| + |e_i^*|} \quad (3)$$

and e_i^* is the *i*th error from a benchmark model

Both the *CumRAE* and the *UMBRAE* error measures are relative to benchmark errors. The benchmark used in this paper is a naïve model that forecasts that out-of-sample temperature anomalies will be the same as the median of the estimation sample anomalies, as described above. In both cases a figure of 1.0 indicates that the errors are equivalent to the benchmark model's forecast errors. A figure below 1.0 indicates that the errors are smaller, and above that they are higher, such that a figure of 0.9 represents an error reduction of 10 percent relative to the benchmark errors and a figure of 1.1 represents an error increase of 10 percent relative to the benchmark errors.

2.3 Test 2: Reliability of alternative causal models [H4]

To obtain out-of-sample forecasts with which to assess how the predictive validity of the models is affected by the successive addition of extra observations to the estimation samples, the models' parameters were estimated using historical data from 1850 to 1899 to forecast temperature data for the nineteen years 2000 to 2018. The process was repeated, adding each year's observation from 1900 to 1999 to the estimation sample in turn, replicating the process of updating an established model with new observations as they become available. The resulting 101 model estimates were each used to forecast the most recent years of available data—from 2000 to 2018. *MdAEs* were calculated for the 19 forecasts from each model (8) estimation period (101) combination.

3. Findings

Table 2 provides a summary table of the Test 1 results. Figure 1 provides charts of the forecast errors from Test 1 that are summarised in Table 2. The implications of the patterns of forecast errors shown in the Figure 1 charts are examined below in the latter half of section 3.4 Reliability of independent versus IPCC models [H4].⁶

3.1 Predictive validity of causal models versus naïve model [H1]

Forecast errors were larger than the benchmark errors (*UMBRAE*) for the IPCC Anthro models AVL and AVSL estimated with data from 1850 to 1949 and from 1850 to 1969, and for the AVR and AVSR models estimated with data from 1850 to 1899, 1850 to 1949, and 1850 to 1969. The anthropogenic warming models showed predictive validity relative the naïve model (*UMBRAE* less than 1.0) for only three of the eight combinations of forecast variable and estimation sample period. Namely, the AVL and AVSL models estimated with data from 1850 to 1899 and from 1850 to 1999, and the AVR and AVSR models estimated with data from 1850 to 1999 (see Table 2).

By contrast, the errors of the forecasts from the independent solar models (S_{BVL} , S_{HVL} , S_{BVR} , and S_{HVR}) were smaller than those of the benchmark model in all cases. In all but one of the eight cases the solar models provided forecasts that reduced error by at least 10 percent, and typically reduced errors by much larger percentages. The one exception was the S_{BVL} model estimated using data from 1850 to 1899 produced forecast errors with an *UMRAE* of 0.964, an error reduction of 5.6 percent relative to the benchmark errors (see Table 2).

3.2 Predictive validity of independent versus IPCC models [H2]

The *MdAEs* of the forecasts from the models with IPCC's anthropogenic and volcanic series as causal variables (AVL and AVR) and from the models that also included IPCC's solar series (AVSL and AVSR) were greater than 1°C (roughly 2°F) for five of the eight combinations tested

⁶ Figures in this paper were created in Microsoft Excel®.

Table 2: Northern Hemisphere temperature models: Fit, predictive validity, and bias

Model variables and estimation statistics				Forecasts: Number, Bias, Errors					Correlations		
	Putative causal variables†	1850 - year (n)	\bar{R}^2	#	Bias‡	MdAE§	IQR§§	CumRAE*	UMBRAE*	Fit vs. Accuracy**	Bias vs. Error††
All Land	Anthro	-	0.155		-0.44	0.41	0.37	0.674	0.715		
	Anthro	IPCC Solar	1899	119	-0.45	0.45	0.34	0.689	0.760	-0.545	0.970
	-	B2000 Solar	(50)		-0.68	0.53	0.73	1.032	0.964		
	-	H1993 Solar			-0.59	0.53	0.61	0.904	0.832		
	Anthro	-	0.508		1.14	1.01	2.09	1.678	1.160		
	Anthro	IPCC Solar	1949	69	1.13	1.01	2.08	1.670	1.155	-0.605	0.997
	-	B2000 Solar	(100)		-0.40	0.35	0.57	0.572	0.527		
	-	H1993 Solar			-0.51	0.43	0.65	0.721	0.679		
	Anthro	-	0.477		1.80	1.76	1.80	2.116	1.809		
	Anthro	IPCC Solar	1969	49	1.69	1.64	1.68	1.983	1.716	-0.850	1.000
	-	B2000 Solar	(120)		-0.44	0.39	0.50	0.528	0.524		
	-	H1993 Solar			-0.59	0.52	0.65	0.703	0.669		
Anthro	-	0.601		0.24	0.22	0.32	0.203	0.193			
Anthro	IPCC Solar	1999	19	0.11	0.18	0.33	0.154	0.152	0.924	0.995	
-	B2000 Solar	(150)		-0.58	0.52	0.30	0.469	0.444			
-	H1993 Solar			-0.88	0.83	0.45	0.706	0.679			
Rural Land	Anthro	-	0.163		-5.08	1.96	7.68	11.415	4.358		
	Anthro	IPCC Solar	1899	119	-5.06	1.86	7.70	11.395	4.498	-0.978	0.999
	-	B2000 Solar	(50)		-0.26	0.27	0.47	0.719	0.785		
	-	H1993 Solar			0.03	0.22	0.42	0.608	0.694		
	Anthro	-	0.264		1.94	1.71	2.99	4.432	3.011		
	Anthro	IPCC Solar	1949	69	1.97	1.73	3.04	4.492	3.003	0.506	1.000
	-	B2000 Solar	(100)		0.17	0.29	0.39	0.672	0.883		
	-	H1993 Solar			-0.11	0.24	0.50	0.723	0.840		
	Anthro	-	0.243		2.28	2.20	1.98	4.116	3.932		
	Anthro	IPCC Solar	1969	49	2.49	2.43	2.20	4.499	4.198	0.029	0.999
	-	B2000 Solar	(120)		-0.02	0.22	0.40	0.512	0.647		
	-	H1993 Solar			-0.20	0.25	0.57	0.666	0.731		
Anthro	-	0.193		0.02	0.22	0.47	0.267	0.272			
Anthro	IPCC Solar	1999	19	-0.04	0.23	0.53	0.269	0.261	-0.914	0.927	
-	B2000 Solar	(150)		-0.39	0.30	0.57	0.449	0.353			
-	H1993 Solar			-0.59	0.54	0.61	0.662	0.576			

† All models were estimated using ordinary least squares regression in STATA and include the variable "Volcanic".

‡ Mean signed error (forecast minus actual, °C).

§ §§ °C. MdAE is median absolute error. IQR is interquartile range calculated from signed errors.

* Cumulative Relative Absolute Error (Armstrong & Collopy 1992) and Unscaled Mean Bounded Absolute Error (Chen, Twycross, & Garibaldi 2017) figures are both relative to a naïve model forecast equal to the median value of the estimation data (a "no-change" forecast). Values of less than 1.0 represent error reductions relative to the naïve method (e.g., 0.95 represents an error reduction of 5%). Conversely values greater than 1.0 represent error increases (e.g., 1.20 represents an error increase of 20%).

** Sign-reversed Pearson correlation between the \bar{R}^2 s and the UMBRAEs.

†† Pearson correlation between the absolute values of Bias (°C) and the UMBRAEs.

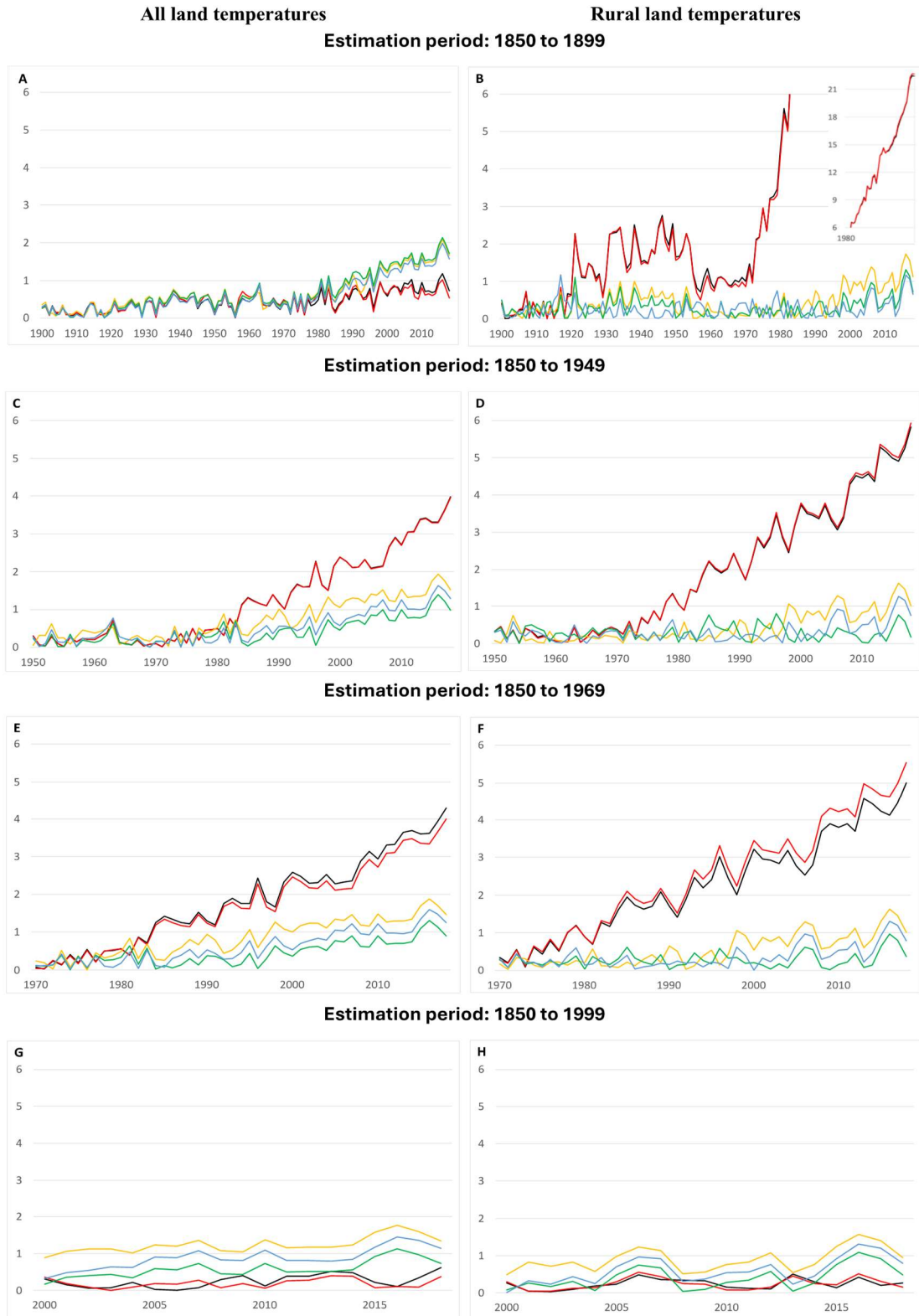


Figure 1: Absolute Errors of NH All Land and Rural Land Temperature Forecasts to 2018 (°C) -- Forecasts from four alternative models plus naïve estimates over four periods.

Legend (Causal variables in models): ■ Anthro, Volcanic; ■ Anthro, Volcanic, IPCC Solar; ■ B2000 Solar, Volcanic; ■ H1993 Solar, Volcanic; ■ Estimation sample median temperature.

(see black and red figures in Table 2). The *MdAEs* of the forecasts from the models with B2000 solar and the volcanic series as causal variables (S_{BVL} and S_{BVR}) were less than 0.55°C (1°F) for all eight of the estimation periods used and temperature series being forecast combinations (see green figures in Table 2), and for seven of the eight in the case of the models with H1993 as the solar variable (S_{HVL} and S_{HVR}) (see blue figures in Table 2).

When estimated from observations to 1969 and used to forecast the subsequent years—which are identified most strongly with the IPCC’s manmade global warming claim—the independent solar models reduced forecast errors (*UMBRAEs*) relative to the IPCC Solar models (AVSL and AVSR) by 69 percent (S_{BVL}), 61 percent (S_{HVL}), 85 percent (S_{BVR}), and 83 percent (S_{HVR}) (from ratios of green and blue figures to red figures in Table 2).

The interquartile ranges (*IQRs*) of the signed errors for all the anthropogenic models tested (AVL, AVSL, AVR, and AVSR) were greater than 1°C for five of the eight combinations of model and estimation period, and greater than 2°C for four of the eight. In contrast, the maximum *IQRs* for the independent models were all substantially less than 1°C , at 0.73°C (S_{BVL}), 0.65°C (S_{HVL}), 0.57°C (S_{BVR}), and 0.61°C (S_{HVR}) (see Table 2).

While predictive validity is concerned with minimising out-of-sample forecast errors, in cases where average absolute forecast errors from alternative models are similar, decision makers might prefer the model that provides forecasts that are less biased—i.e., that have a smaller average signed error. In this study, however, there is no trade-off between forecast accuracy and bias: Absolute forecast errors (*UMBRAE*) were strongly correlated with forecast bias (absolute value of mean signed errors) for all eight combinations of causal model and estimation period: the more the bias, the larger the error. Pearson’s *r* values ranged from 0.93 to 1.00 (see the rightmost column of Table 2).

Average signed errors ranged between -0.9°C and 1.8°C for the All Land models, and between -5.1°C and 2.5°C for the Rural Land models (see the Bias column of Table 2). The errors of forecasts from the anthropogenic models for the era of concern over manmade global warming, starting in 1970, were 1.8°C (AVL), 1.7°C (AVSL), 2.3°C (AVR), and 2.5°C (AVSR) warmer than the measured temperatures. On the other hand, the errors of forecasts from the independent solar models were 0.4°C (S_{BVL}), 0.6°C (S_{HVL}), 0.0°C (S_{BVR}), and 0.2°C (S_{HVR}) cooler than the measured temperatures.

3.3 Relationship between predictive validity and statistical fit of models [H3]

The correlations (sign-reversed Pearson’s *r*) between the accuracy of out-of-sample forecasts, as measured by *UMBRAE* (an error measure, hence the sign reversal), and the statistical fit of the models to the estimation data (adjusted- R^2) for the causal models tested were large and *negative* for five (5) of the eight (8) combinations of estimation period (1850 to 1899, 1949, 1969, and 1999) used—and hence maximum forecast horizon of 119, 69, 49, and 19 years, respectively—and temperature series (NH Land and NH Rural) forecast (see “Fit vs. Accuracy” column of Table 2).

The average of the individual correlations was -0.30 , indicating that better statistical fit—i.e., larger adjusted- R^2 —was, on average, associated with *lesser* forecast accuracy (higher *UMBRAE*). Only for forecasts of NH Land from models estimated with data from 1850 to 1999—the largest estimation sample size used for this testing—was there a substantial positive correlation between statistical fit and forecast accuracy.

3.4 Reliability of independent versus IPCC models [H4]

Charts of the results of Test 2 are presented in Figure 2 and are discussed below.

The independent solar models— S_{BVL} and S_{HVL} , and S_{BVR} and S_{HVR} —perform largely as one would expect of causal models when forecasting using known values of the causal variables.

Firstly, the *MdAEs* of the forecasts for the years 2000 to 2018 are less than those of the naïve benchmark model for estimation samples of 50 observations (1850 to 1899) and more, with two minor exceptions. In particular, the S_{BVL} model when estimated with the 33 samples 1850 to 1899 through to 1850 to 1932 (observation 83) produced *MdAEs* greater than the naïve model. At worst, the S_{BVL} model produced a *MdAE* 30 percent greater than the naïve model when it was estimated from the first 66 historical observations (1850 to 1915). And the S_{BVR} model produced forecasts with an *MdAE* greater than that of the corresponding naïve model in the case of only one of the 101 estimation samples tested. The S_{HVL} and S_{HVR} models—each estimated over 101 samples—provided forecasts with *MdAEs* that were smaller than the corresponding Naïve model *MdAEs* for all the estimation samples.

Secondly, the *MdAEs* of the independent solar model forecasts trend broadly and moderately downwards as additional observations are added to the estimation samples. And, thirdly, the *MdAEs* of the S_{BVR} and S_{HVR} models were 66 percent and 71 percent smaller, respectively, than those of the S_{BVL} and S_{HVL} models when the models were estimated using the first 50 observations and 42 percent and 35 percent smaller when estimated using all 150 observations up to 1999. In other words, the independent solar models provided forecasts that were more accurate when forecasting temperature anomalies that excluded urban heat island effects.

The performances of the IPCC inspired models—AVL, AVSL, AVR, and AVSR—especially when the models are applied to forecasting rural only temperatures—AVR and AVSR—were markedly different. In the latter cases, forecasts for the years 2000 to 2019 from models estimated with 50 observations of historical data (1850 to 1899) have *MdAEs* of around 17 °C or 1600 percent greater than the 1 °C *MdAE* of forecasts from the naïve benchmark model.

As Figure 2 shows, the *MdAEs* of the IPCC inspired model forecasts are greater, mostly much greater, than the those of the naïve model forecasts for most estimation sample sizes. The *MdAEs* of the AVL and AVSL model forecasts were greater than the naïve model forecasts for 64 and 63, respectively, of the 101 sample sizes over which the models were estimated. For the AVR and AVSR models—estimated to forecast rural only temperatures—the figures were both 79.

Puzzlingly, the *MdAE* of the forecasts from the AVSR model, when estimated using all 150 observations up to 1999, was 29 percent *larger* than the corresponding *MdAE* of forecasts from the AVSL model. Moreover, the *MdAE* of the AVSR model forecasts was 5 percent greater than that of the AVR's. In other words, the inclusion of the IPCC Solar variable *reduced* the model's predictive validity for forecasting temperatures unaffected by urban heat islands when it was estimated using the largest sample of data (150 observations) employed in this study.

The *MdAEs* of the IPCC affiliated model forecasts varied wildly up and down as more observations were added to the estimation samples. After oscillating with each additional observation added to the estimation sample from 1899 to 1921 (50 to 72 observations) *MdAEs* of the forecasts from the AVL model broadly increased as additional observations were added until the observations for 1974 were added (125 observations in the sample), at which point the *MdAEs* began to decline with each additional observation. The AVSL model forecast *MdAEs* followed a closely similar pattern.

In the case of the AVR and AVSR models—forecasting the rural land temperatures, on the right of Figure 2—the *MdAEs* decreased rapidly from roughly 17 times the corresponding naïve forecast errors to beat the naïve *MdAE* when the 76th observation (1925) was added to the estimation samples. After that observation was added, the *MdAEs* for the AVR and AVSR model forecasts increased rapidly with each extra observation then stayed high before rapidly declining again after the 116th observation (1965) was added to the estimation samples.

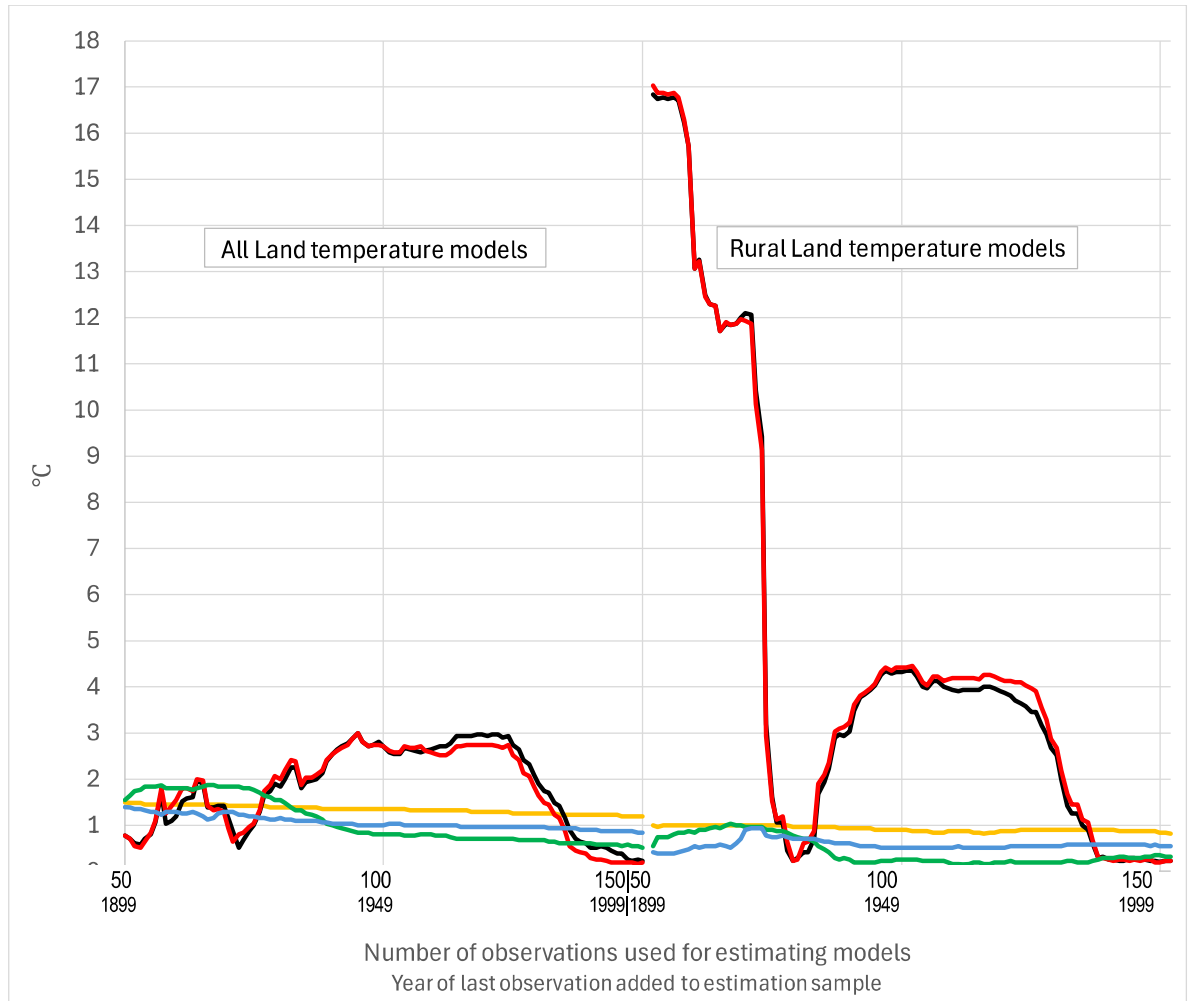


Figure 2. Median absolute errors of NH temperature forecasts 2000 to 2018 in °C. Legend (Causal variables in models): ■Anthro, Volcanic; ■Anthro, Volcanic, IPCC Solar; ■B2000 Solar, Volcanic; ■H1993 Solar, Volcanic; ■Estimation sample median temperature.

In the case of the AVR and AVSR models—forecasting the rural land temperatures, on the right of Figure 2—the *MdAEs* decreased rapidly from roughly 17 times the corresponding naïve forecast errors to beat the naïve *MdAE* when the 76th observation (1925) was added to the estimation samples. After that observation was added, the *MdAEs* for the AVR and AVSR model forecasts increased rapidly with each extra observation then stayed high before rapidly declining again after the 116th observation (1965) was added to the estimation samples.

When a model of causal relationships is estimated from empirical data on valid causal variables reliably measured, one would expect forecast errors to get smaller as more observations are used in the estimation of the model’s parameters. That is what the charts in Figure 2 show in the case of the naïve benchmark model forecasts and, broadly, what can be seen in the case of the independent models S_BVL , S_HVL , S_BVR , and S_HVR , but is not seen in the case of the models using the IPCC variables: AVL , $AVSL$, AVR , and $AVSR$.

The median absolute errors of the models’ forecasts in Figure 2 can be modelled as linear functions of the number of observations in the estimation samples. The residuals—within-sample forecast errors—of OLS-estimated *MdAE*-prediction-models have narrow 5th to 95th inter-percentile ranges of -0.015 to 0.017 °C and -0.065 to 0.038 °C for the benchmark models of All Land and

of Rural Land temperatures, respectively. In the case of the independent models, the equivalent inter-percentile ranges were -0.303 to 0.290 °C (S_{BVL}), -0.069 to 0.075 °C (S_{HVL}), -0.285 to 0.282 °C (S_{BVR}), and -0.186 to 0.170 °C (S_{HVR}). None were more than 0.6 °C between the 5th and 95th percentiles.

The equivalent inter-percentile ranges for the IPCC Anthro-variable models' residuals were considerably larger, each covering several whole degrees Celsius, with none smaller than 2.46 °C. The ranges were: -1.45 to 1.13 °C (AVL), -1.41 to 1.05 °C ($AVSL$), -7.02 to 5.60 °C (AVR), and -6.96 to 5.70 °C ($AVSR$).

The charts in Figure 1, above, give another perspective on the relationships between the number of observations in model estimation samples and the errors of the model's forecasts. From top to bottom of Figure 1, the four pairs of charts show the absolute errors of out-of-sample forecasts - albeit using known values of causal variables except for the benchmarks - from models estimated from four increasingly large samples of data. The forecasts spanned the out-of-sample periods from the years 1900, 1950, 1970, and 2000 to 2018.

What should those forecast error charts look like? First, because the models are causal - without time or lagged variables - and "known" values of the causal variables are used for forecasting, errors should only increase over longer forecast horizons to the extent that the out-of-sample period includes observations that are novel relative to the estimation sample. The likelihood of novelty should decrease with increasing estimation sample size.

Second, and related to the first point, forecast errors for the same horizon should diminish with increasing estimation sample size, to the point that the errors reflect only random variation or the effects of missing or mis-specified or mal-estimated variables. One would expect little or no increase in errors for longer horizons when the estimation sample is large and representative, and the model and data realistically represent of the situation.

The absolute forecast errors from the benchmark models and from the independent solar models shown in the Figure 1 charts— S_{BVL} , S_{HVL} , S_{BVR} , and S_{HVR} —seldom exceed 1 °C and follow the paths expected of well-specified models, albeit to a lesser extent in the case of the models forecasting the heat-island-affected All Land temperatures. The IPCC Anthro models— AVL , $AVSL$, AVR , and $AVSR$ —do not.

The errors of the Anthro models' forecast errors explode well beyond 1 °C and the benchmark model errors for forecast years beyond the mid-1970s, with puzzling exceptions. Namely, forecasts from Anthro models estimated from the largest sample size in the chart—1850 to 1999—and from models estimated from the smallest sample—1850 to 1899—forecasting All Land temperatures. In those cases, involving three of the eight charts, the Anthro model errors are less than the median historical temperature benchmark model errors, and mostly less than the errors of the independent models in later years.

The explosion in Anthro model errors from the 1970s is more extreme for models estimated to forecast Rural Land temperatures. Moreover, for the models estimated using only 1850 to 1899 data, errors are larger than those of the benchmark and independent models from 1920 and, prior to 1970, without any obvious pattern.

4. Discussion

The findings of this study that relate to the forecasting of the independent estimate of Northern Hemisphere rural land temperatures are most apposite to answering the IPCC questions paraphrased here as, "what has caused changes in annual average temperatures on Earth over recent decades and what, therefore, can we expect in the way of temperature changes over coming decades?" The NH Rural Land temperature series was developed in order provide a large area annual average estimate of surface temperatures that would reflect changes in the global climate uncontaminated by local non-climate influences on measurement, specifically heat island effects.

Based on the models estimated to forecast NH Rural Land temperatures, then, the four hypotheses tested in this study were largely affirmed in their expected, [rather than alternate], directions when known estimates of the causal variable values were used for forecasting the unknown-to-the-model temperature variable.

The findings on the predictive validity [H1] of the IPCC Anthro models were the only, and partial, exception. The cumulative absolute errors of out-of-sample forecasts from models estimated using samples from 1850 to 1899, to 1949, and to 1969 were, on average, nearly twelve times greater than the benchmark model errors in the first case and more than four times greater in the latter two. Only forecast errors from models estimated using data from 1850 to 1999 to forecast temperatures for the years 2000 to 2018 were smaller than the benchmark model errors and, remarkably, smaller than those of the independent solar models (see Table 2 and Figure 1).

The findings of this study beg the question: Why did the IPCC anthropogenic models provide forecasts that were so grossly inaccurate in absolute terms, relative to a naïve benchmark model based only on historical data on the temperature variable to be forecast, and relative to independent solar causal models?

We suggest that the broad answer is that the IPCC was established by government officials with the objective of finding substantive human influence on global temperatures⁷ rather than to discover useful knowledge on climate change by testing plausible alternative hypotheses developed from prior knowledge. Hence this study's H2 hypothesis. Armstrong and Green (2022) refer to the antiscientific practice of undertaking research designed to support a given hypothesis as “advocacy research,” a practice unlikely to produce useful knowledge and that risks harm through unnecessary worry and bad personal and policy decisions.

The independent solar models were more reliable than the IPCC models [H4] and provided forecasts that were more accurate than those from the naïve benchmark and, in most tests, than those from the IPCC models [H2]. The independent models performed as expected of models of causal relationships that were developed using prior knowledge and established theory and with parameters that were estimated using data that represented the causal and forecast variables realistically. Out-of-sample forecast errors were consistently smaller than those of a simple benchmark model and generally declined with larger estimation samples. The IPCC anthropogenic climate change models did *not* perform as one would expect of valid models estimated using valid and reliable data (see Figure 2).

The great variations, mostly large size, and lack of expected pattern in IPCC model forecast errors as additional observations are added to the estimation sample, are puzzling. While the extremes are not so pronounced with the forecasts of NH All Land temperatures as they are with the NH Rural Land temperatures, the strangeness of forecast errors increasing dramatically when additional observations are added is common to both.

While the independent solar models provided valid and reliable forecasts, the model that included the IPCC solar variable along with the IPCC's Anthro and Volcanic variables did not. The model's forecast errors were, for the most part, almost identical to those from the IPCC model that did not include the IPCC solar variable. The exceptions are enigmatic. When models were estimated using data from 1850 to 1969, a gap begins to open between the errors of forecasts from the model including the IPCC solar variable and the model that did not for forecasts for the early 1980s and beyond (Figure 1.) The gap is more pronounced in the forecasts of the NH Rural Land

⁷ See for example, UN IPCC's “About” page (<https://www.ipcc.ch/about/>) where it is stated that “The IPCC provides regular assessments of the scientific basis of climate change, its impacts and future risks, and options for adaptation and mitigation. ... Created in 1988 by the World Meteorological Organization (WMO) and the United Nations Environment Programme (UNEP), the objective of the IPCC is to provide governments at all levels with scientific information that they use to develop climate policies. The IPCC reports are also a key input into international climate change negotiations.”

temperatures than it is in the All-Land forecasts. Moreover, the relative sizes of the forecast errors are reversed between the Rural and All-Land temperatures.

4.1 Forecasting test extension

The tests the present paper describes were all intended to provide realistic representations of the forecasting problem in the sense that the models were estimated using observations from earlier in time and were then used to forecast temperatures later in time. Because the models are causal and do not use time or lagged-dependent variables, however, one would expect to get similar results—out-of-sample forecast errors—from models estimated using any *representative* subsample of the data. For example, *if* the estimates of variable values for the 84 years from 1850 to 1933 include sufficient diversity to allow realistic estimation of the causal relationships, then the forecast errors for the 85 years from 1934 to 2018 should on average be like those for 85 forecasts for even-numbered years from models estimated from odd years data.

Table 3: Forecast errors from models estimated with data from 1850 to 1933 versus those from models estimated with data from odd-numbered years

	NH All Land			NH Rural Land		
	Bias [†]	MdAE [‡]	IQR [§]	Bias [†]	MdAE [‡]	IQR [§]
Estimation sample: 84 consecutive years from 1850 to 1933						
Anthro, Volcanic	0.544	0.34	1.29	0.705	0.56	1.53
Anthro, Volcanic, IPCC Solar	0.531	0.37	1.31	0.685	0.56	1.57
B2000 Solar, Volcanic	-0.685	0.53	0.68	-0.173	0.23	0.50
H1993 Solar, Volcanic	-0.510	0.36	0.60	-0.247	0.25	0.51
Estimation sample: 84 odd-numbered years from 1851 to 2017						
Anthro, Volcanic	0.033	0.15	0.32	0.005	0.20	0.42
Anthro, Volcanic, IPCC Solar	0.034	0.15	0.30	0.005	0.22	0.43
B2000 Solar, Volcanic	0.029	0.18	0.34	0.002	0.23	0.42
H1993 Solar, Volcanic	0.025	0.22	0.40	-0.001	0.22	0.48
Ratios of error statistics (even-years' / 1934 to 2018 forecasts)						
Anthro, Volcanic	0.061	0.43	0.25	0.007	0.36	0.28
Anthro, Volcanic, IPCC Solar	0.064	0.41	0.23	0.007	0.39	0.28
B2000 Solar, Volcanic	0.042	0.35	0.50	0.009	0.99	0.85
H1993 Solar, Volcanic	0.050	0.61	0.66	0.005	0.89	0.96

[†] Mean signed error (forecast minus actual, °C).

^{‡ §} Both are °C. **MdAE** is median absolute error. **IQR** is interquartile range and is calculated from signed errors.

Should the forecast errors from such a comparison be dissimilar, the implications are that the earlier data are not adequately representative, or the model is unrealistic, or the estimates of the values of the variables are invalid or unreliable, or all or some combination of the foregoing. Regarding the first point, given that climate changes unfold slowly on human timescales, one would expect that models estimated using every second year's data would, *to some extent*, better represent the causal relationships than would models estimated from only the earlier half of the available 169 years of data.

Table 3 goes some way to disentangling the effects. The bolded figures for NH Rural Land *MdAEs* and *IQRs* for the independent models show little or no reduction from estimating models using every second year from the full period of the available data. Those figures contrast with the figures

for the NH All Land forecast errors from the IPCC-inspired models, which show reductions in *MdAE* of 57 percent and 59 percent and reductions in the *IQRs* of errors of 75 and 77 percent.

Those results suggest that the independent models of NH Rural Land temperatures and the estimates of the variable values provide realistic representations of causal relationships, and that the observations for the first 84 years of available data are sufficiently representative to provide useful estimates of the relationships. The same cannot be said of the IPCC-inspired models. Notably, the IPCC solar models have small negative coefficients for the solar variable when estimated with 1850 to 1933 data, but small positive ones when estimated using data from odd-numbered years. And the coefficient for the Anthro variable estimated using data from odd-numbered years is less than a half of what it is when estimated with 1850 to 1933 data. The opposite is the case with the independent solar models, with the solar coefficients increasing by more than half when estimated using the likely more-representative odd-numbered years data (see Table 4).

Table 4: Parameters of models estimated with standardised data from 1850 to 1933 versus those from models estimated with standardised data from odd-numbered years

Model variables		Parameters (estimation data used & ratio)		
Forecast	Causal	1850-1933	Odd Years	Ratio
All Land (IPCC)	Volcanic	0.08	0.03	0.35
	Anthro	1.13	0.45	0.40
	IPCC Solar	-0.04	0.07	-1.98
Rural Land (Independent)	Volcanic	0.12	0.11	0.95
	H1993 Solar	0.17	0.31	1.88
	Volcanic	0.13	0.15	0.95
	B2000 Solar	0.20	0.33	1.67

None of the above inspires confidence in the IPCC models. If models estimated using the first half of the available data perform relatively poorly in forecasting the second half using “known” values of the putative causal variables, why should policy makers take account of the models’ forecasts for decades to come when alternative independent models without an anthropogenic causal variable are shown to be more reliable and to provide forecasts that are on average more accurate?

An important implication of the lack of predictive validity of the IPCC anthropogenic models and the validity of the non-anthropogenic models, is that human emissions of carbon dioxide, etc., have had little or no influence on NH temperatures and cannot therefore be regarded as a policy variable. The independent causal models of NH temperature tested in this study and found to have predictive validity did not include an anthropogenic emissions variable. Each included an independently developed solar irradiance variable and the IPCC’s Volcanic variable. Neither solar irradiance nor volcanic emissions are usefully predictable given the current state of knowledge. Nor are they policy variables, given they are not subject to policymaker control on a global scale.

4.2 Further research

This study tested causal models that included two of the 27 independent estimates of total solar irradiance identified by Connolly et al. (2023). Testing the predictive validity of models incorporating others of the solar variables that are well supported by theory and evidence may identify better estimates of TSI than those used in this study and help researcher to develop still better estimates.

Neither this study, nor Connolly, et al.'s (2023) study, considered alternatives to the IPCC's Volcanic estimates. Perhaps improvements could be made in estimating the values of that important causal variable with the aim of further improving the realism and hence predictive validity of independent solar models.

The non-causal benchmark model used in this study—the median estimation sample temperature was extrapolated as the forecast of all future temperatures—provided forecasts that were mostly more accurate than those from the IPCC models and that were typically within 1 °C of the estimated NH Rural Land temperature series being forecast. The remarkable success of that simple model suggests the possibility that a more sophisticated extrapolation model might be competitive with the independent causal models, especially when the difficulty of forecasting the causal variables is considered.

One extrapolation method that has been found to reduce forecasts errors in many and diverse tests—and could be tested for the temperature forecasting problem—is damped-trend exponential smoothing (Gardner 2006). A simple two-observation variation of the damped trend method was used to forecast week-ahead doctor visits for influenza. The model's forecasts reduced errors over a 440-week period by 54 percent relative the complex and expensive Google Flu Trends machine learning model.⁸

A key finding from decades of forecasting research is that combining forecasts from diverse *valid* methods and models reduces forecast errors on average by as much as one-half (Armstrong and Green, 2018). To the extent that further research reveals more methods, models, and data that are valid for the temperature forecasting problem, incorporating more knowledge in forecasts by combining the forecasts is likely to reduce errors.

Finally, perhaps it will be possible to forecast the solar and volcanic variables to the extent that the resulting out-of-sample causal model forecast errors are less than those of the best historical temperature extrapolation model.⁹ From a practical point of view, however, it seems unlikely that even large reductions in forecast errors would have policy or planning implications. Research on the level of uncertainty—empirical prediction intervals—consistent with using the forecasts for policy or planning is needed.

5. Conclusions

The IPCC's models of anthropogenic climate change lack predictive validity. The IPCC models' forecast errors were greater for most estimation samples—often many times greater—than those from a benchmark model that simply predicts that future years' temperatures will be the same as the historical median. The size of the forecast errors and unreliability of the models' forecasts in response to additional observations in the estimation sample implies that the anthropogenic models fail to realistically capture and represent the causes of Earth's surface temperature changes. In practice, the IPCC models' relative forecast errors would be still greater due to the uncertainty in forecasting the models' causal variables, particularly Volcanic and IPCC Solar.

The independent solar models of climate change—which did not include a variable representing the IPCC postulated anthropogenic influence—*do* have predictive validity. The models reduced errors of forecasts for the years 2000 to 2018 relative to the benchmark errors for all, and all-but-one, of 101 estimation samples tested for each of the two models. One of the models (B2000

⁸ A write-up of the study, titled "Forecasts of doctor visits for flu: Simple conservative methods beat Google's big data machine learning model", is available on *ResearchGate* at <http://dx.doi.org/10.13140/RG.2.2.25199.56487/1>

⁹ For example, see Velasco Herrera, *et al.*'s (2021) attempts to forecast the so-called sunspot activity variations. Sunspot activity has non-trivial connections to solar irradiance variations, as was described in Connolly *et al.* (2021, 2023).

Solar) reduced errors by more than 75 percent for forecasts from models estimated from 35 of the samples—a particularly impressive improvement given that the benchmark errors were no greater than 1 °C for all but one of the estimation samples.

The independent solar models provide realistic representations of the causal relationships with surface temperatures. The question of whether the independent solar variables can be forecast with sufficient accuracy to improve on the benchmark model forecasts in practice, however, remains relevant. All in all, and contra to the IPCC reports, there is insufficient evidential basis for the use of carbon dioxide, et cetera, emissions—taken together, the IPCC’s Anthro—as climate policy variables.

Finally, this study provides further evidence that measures of statistical fit provide *misinformation* about predictive validity. Predictive validity can only be properly estimated when the proposed model or hypothesis is used for forecasting new-to-the-model data, and the forecasts are then compared for accuracy against forecasts from a plausible benchmark model. This important conclusion needs bearing-in-mind when evaluating policy models.

Funding

The authors did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors for the research presented in this paper.

Co-Editor: Stein Storlie Bergsmark

Reviewers: Anonymous

Other Statements and Declarations

The authors declare that they have no competing interests.

Acknowledgments

John Dawes, Gerd Gigerenzer, Paul Myers, Keith Ord, Darrell Velegol, and Arch Woodside made helpful suggestions on drafts of the paper. Four anonymous reviewers critiqued the submitted paper. A presentation of a draft of the paper to colleagues of the first author’s school provided positive feedback.

Data Availability Statement

The data used in this study are provided by Connolly (2023) and are available at <https://zenodo.org/records/8225275> (DOI 10.5281/zenodo.8225274).

References

- Allen, M.R. & Tett, S.F.B. (1999). Checking for model consistency in optimal fingerprinting. *Climate Dynamics*, 15, 419-434. <https://doi.org/10.1007/s003820050291>
- Armstrong J.S. (2001). Evaluating forecasting methods. In: Armstrong J.S. (Ed.) *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Kluwer, Norwell MA, pp 443-472. https://doi.org/10.1007/978-0-306-47630-3_20
- Armstrong, J.S. & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, 8, 69-80. [https://doi.org/10.1016/0169-2070\(92\)90008-W](https://doi.org/10.1016/0169-2070(92)90008-W)
- Armstrong, J.S. & Green, K.C. (2018). Forecasting methods and principles: Evidence-based checklists. *Journal of Global Scholars of Marketing Science*, 28, 103-159. <https://doi.org/10.1080/21639159.2018.1441735>

- Armstrong, J.S. & Green, K.C. (2022). *The Scientific Method: A Guide to Finding Useful Knowledge*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781009092265>
- Armstrong, J.S., Green, K.C. & Graefe, A. (2015). Golden rule of forecasting: Be conservative. *Journal of Business Research*, 68, 1717-1731. <https://doi.org/10.1016/j.jbusres.2015.03.031>
- Bard, E., Raisbeck, G., Yiou, F. & Jouzel, J. (2000). Solar irradiance during the last 1200 years based on cosmogenic nuclides. *Tellus B*, 52, 985-992. <http://dx.doi.org/10.1034/j.1600-0889.2000.d01-7.x>
- Bailey, R. (2000). Earth Day, then and now: The planet's future has never looked better. Here's why. *Reason*, May. <https://reason.com/2000/05/01/earth-day-then-and-now-2/>
- Beverley, J.D., Newman, M. & Hoell, A. (2024). Climate model trend errors are evident in seasonal forecasts at short leads. *Climate and Atmospheric Science*, 7, 285. <https://doi.org/10.1038/s41612-024-00832-w>
- Chen, C., Twycross, J. & Garibaldi, J.M. (2017). A new accuracy measure based on bounded relative error for time series forecasting. *PLoS ONE*, 12, e0174202. <https://doi.org/10.1371/journal.pone.0174202>
- [dataset] Connolly, R. (2023). Supplementary Materials for "Challenges in the detection and attribution of Northern Hemisphere surface temperature trends since 1850". Zenodo, August 18. Available at <https://doi.org/10.5281/zenodo.8225275>
- Connolly, R., Soon, W., Connolly, M., Baliunas, S., Berglund, J., Butler, C.J., Cionco, R.G., Elias, A.G., Fedorov, V.M., Harde, H., Henry, G.W., Hoyt, D.V., Humlum, O., Legates, D.R., Lüning, S., Scafetta, N., Solheim, J.-E., Szarka, L., van Loon, H., Velasco Herrera, V.M., Willson, R.C., Yan, H. & Zhang, W. (2021). How much has the Sun influenced Northern Hemisphere temperature trends? An ongoing debate. *Research in Astronomy and Astrophysics*, 21, 131. <https://iopscience.iop.org/article/10.1088/1674-4527/21/6/131>
- Connolly, R., Soon, W., Connolly, M., Baliunas, S., Berglund, J., Butler, C.J., Cionco, R.G., Elias, A.G., Fedorov, V.M., Harde, H., Henry, G.W., Hoyt, D.V., Humlum, O., Legates, D.R., Scafetta, N., Solheim, J.-E., Szarka, L., Velasco Herrera, V.M., Yan, H. & Zhang, W. (2023). Challenges in the Detection and Attribution of Northern Hemisphere Surface Temperature Trends Since 1850. *Research in Astronomy and Astrophysics*, 23, 105015. <https://iopscience.iop.org/article/10.1088/1674-4527/acf18e>
- Eyring, V., Cox, P.M., Flato, G.M., et. al. (2019). Taking climate model evaluation to the next level. *Nature Climate Change*. 9, 102–110. <https://doi.org/10.1038/s41558-018-0355-y>
- Gardner, E.S. Jr. (2006). Exponential smoothing: The state of the art – Part II. *International Journal of Forecasting*, 22, 637–677. <https://doi.org/10.1016/j.ijforecast.2006.03.005>
- Green, K.C. & Armstrong, J.S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research*, 68, 1678-1685. <https://doi.org/10.1016/j.jbusres.2015.03.026>
- Green, K.C., Armstrong, J.S. & Soon W. (2009). Validity of climate change forecasting for public policy decision making. *International Journal of Forecasting*, 25, 826-832. <https://doi.org/10.1016/j.ijforecast.2009.05.011>
- Hasselmann, K., Bengtsson, L., Cubasch, U., Hegerl, G.C., Rodhe, H., Roeckner, E., et al. (1995). Detection of anthropogenic climate change using a fingerprint method. In: P. Ditlevsen, (Ed.) *Modern dynamical meteorology: Proceedings from a symposium in honor of Prof. Aksel Wiin-Nielsen*. Copenhagen: University of Copenhagen. Department of Geophysics, pp. 203-221. <https://hdl.handle.net/21.11116/0000-0000-3F03-7>
- Hegerl, G.C., Hasselmann, K., Cubasch, U., Mitchell, J.F.B., Roeckner, E., Voss, R. & Waszkewitz, J. (1997). Multi-fingerprint detection and attribution analysis of greenhouse gas,

greenhouse gas-plus-aerosol and solar forced climate change. *Climate Dynamics*, 13, 613-634. <https://doi.org/10.1007/s003820050186>

Historian (2023). Postal history: Rates for stamped cards and postcards. Unites States Postal Service, <https://about.usps.com/who/profile/history/postcard-rates-since-1873.htm>. Accessed on 8 January 2024.

Hoyt, D.V. & Schatten, K.H. (1993). A discussion of plausible solar irradiance variations, 1700-1992. *Journal of Geophysical Research*, 98, 18895-18906. <https://doi.org/10.1029/93JA01944>

IPCC (1992). What would we now estimate for climate change? In: J.T. Houghton, B.A. Callander & S.K. Varney (Eds.) *Climate change 1992: The supplementary report to the IPCC scientific assessment*. Cambridge: Cambridge University Press, p. 17. https://archive.ipcc.ch/publications_and_data/publications_ipcc_supplementary_report_1992_wg1.shtml

IPCC (2007). *The Physical Science Basis. Working Group I Contribution to the Fourth Assessment Report of the IPCC*. Cambridge: Cambridge University Press. <https://www.ipcc.ch/report/ar4/wg1/>

IPCC (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781009157896>

Matthes, K. et al. (2017). Solar forcing for CMIP6 (v3.2). *Geoscientific Model Development*, 10, 2247-2302. <https://doi.org/10.5194/gmd-10-2247-2017>

McKittrick, R. (2022). Checking for model consistency in optimal fingerprinting: a comment. *Climate Dynamics*, 58, 405-411. <https://doi.org/10.1007/s00382-021-05913-7>

Notz, D. (2015). How well must climate models agree with observations? *Philosophical Transactions of the Royal Society A*, 373, 20140164. <http://dx.doi.org/10.1098/rsta.2014.0164>

Santer, B.D., Mikolajewicz, U., Bruggemann, W., Cubasch U., Hasselmann, K., Hock, H., Maier-Reimer, E. & Wigley T.M.L. (1995). Ocean variability and its influence on the detectability of greenhouse warming signals. *Journal Of Geophysical Research-Oceans*, 100, 10693-10725. <https://doi.org/10.1029/95JC00683>

Santer, B.D., Wigley, T.M.L., Barnett, T.P. & Anyamba, E. (1996). Detection of climate change and attribution of causes. In: J.T. Houghton, L.G. Meira Filho, B.A. Callander, N. Harris, A. Kattenberg & K. Maskell K (Eds.) *Climate change 1995: the science of climate change. Contribution of Working Group I to the Second Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press, 407-443. <https://www.ipcc.ch/report/ar2/wg1/>

Soon, W., Baliunas, S., Idso, S.B., Kondratyev, K.Y. & Posmentier, E.S. (2001). Modeling climatic effects of anthropogenic carbon dioxide emissions: unknowns and uncertainties. *Climate Research*, 18, 259-275. <https://www.int-res.com/articles/cr/18/c018p259.pdf>

Soon, W., Connolly, R. & Connolly, M. (2015). Re-evaluating the role of solar variability on Northern Hemisphere temperature trends since the 19th century. *Earth-Science Reviews*, 150, 409-451. <https://doi.org/10.1016/j.earscirev.2015.08.010>

Soon, W., et al. (2023). The Detection and Attribution of Northern Hemisphere Land Surface Warming (1850-2018) in Terms of Human and Natural Factors: Challenges of Inadequate Data. *Climate*, 11, 179. <https://doi.org/10.3390/cli11090179>

Velasco Herrera, V.M., Soon, W. & Legates, D.R. (2021). Does Machine Learning reconstruct missing sunspots and forecast a new solar minimum? *Advances in Space Research*, 68, 1485-1501. <https://doi.org/10.1016/j.asr.2021.03.023>

Wang, Y., Huang, G., Pan, B., Lin, P., Boers, N., Tao, W., Chen, Y., Liu, B. & Li, H. (2024). Correcting Climate Model Sea Surface Temperature Simulations with Generative Adversarial

Networks: Climatology, Interannual Variability, and Extremes. *Advances in Atmospheric Sciences*, 41, 7, 1299–1312. <https://doi.org/10.1007/s00376-024-3288-6>

Appendix A: Relationships between alternative causal variables and temperature

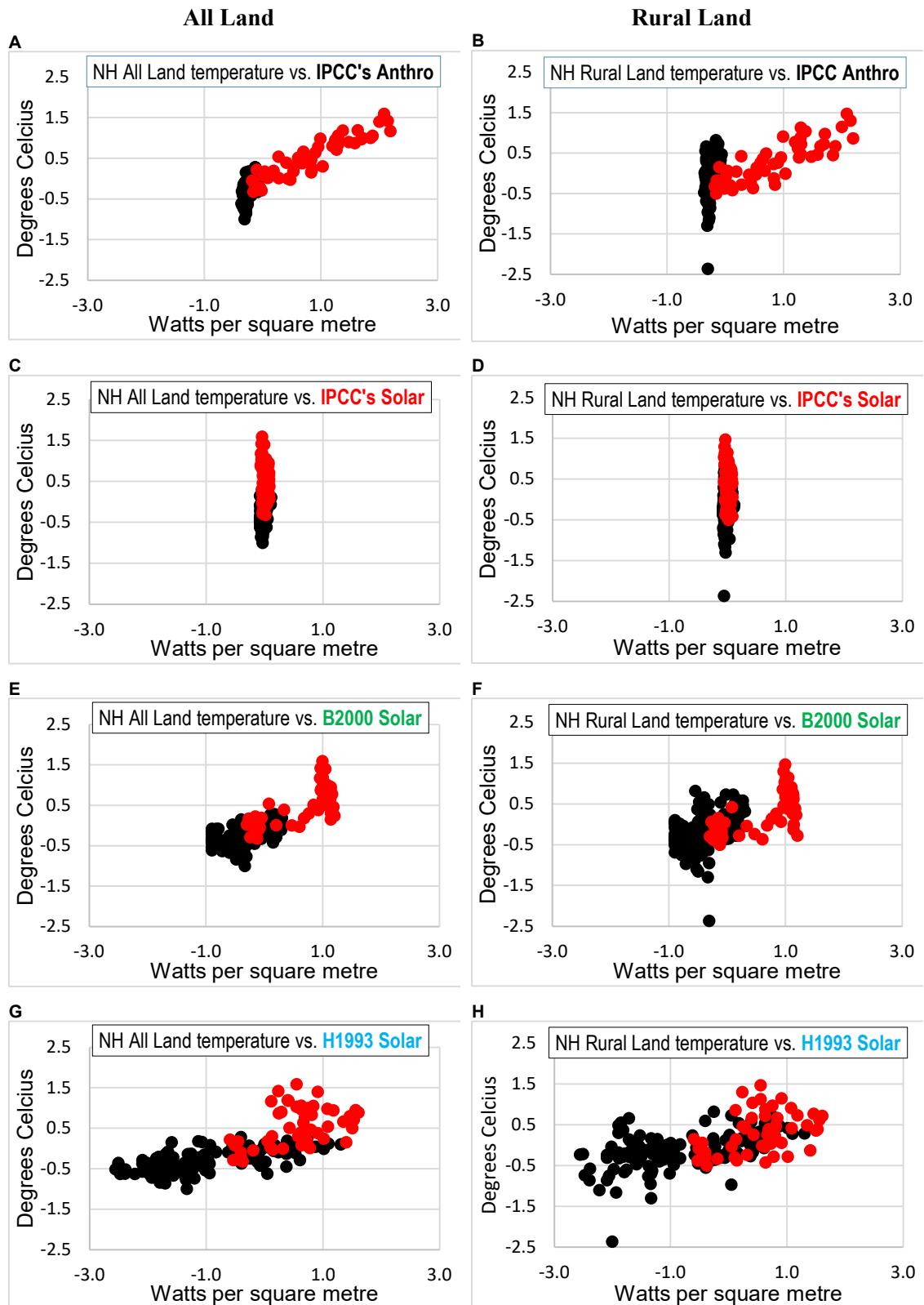


Figure A1: Relationships between Northern Hemisphere All Land and Rural Land Temperature (°C) and alternative causal variables, Legend: ● Observations prior to 1970; ● Observations 1970 to 2018.

Appendix B: Findings on models estimated using stationary causal variables

Causal variables Anthro (CO₂ etc.), Solar IPCC, Solar B2000, and Solar H1993 were transformed to achieve stationarity¹⁰ after conducting augmented Dicky-Fuller unit root tests on all variables on the pre-1900 data by taking first differences (second differences in the case of Solar B2000). The dependent (temperature) variables were stationary, as was the Volcanic variable (all $p = 0.00$). Table B1 highlights the variables that were transformed to achieve stationarity using **white text on black**.

Durbin-Watson tests rejected autocorrelation of errors at the 1 percent level for the models estimated using the stationary data from 1851, or 1852, to 1899; the first (shortest) of the estimation periods.

Table B1: Models and Variables

Model Name	Causal variables			Forecast variable
AVL	Anthro	Volcanic	-	NH All Land Annual Average Temperature Anomaly
AVSL	Anthro	Volcanic	Solar IPCC	
S _B VL	Solar	Volcanic	-	
S _H VL	Solar	Volcanic	-	
AVR	Anthro	Volcanic	-	NH Rural Land Annual Average Temperature Anomaly
AVSR	Anthro	Volcanic	Solar IPCC	
S _B VR	Solar	Volcanic	-	
S _H VR	Solar	Volcanic	-	

The eight models described in Table B1 were estimated using data from 1851 (1852 in the case of the S_BVL and S_BVR models) to 1899, to 1949, to 1969, and to 1999 to provide out-of-sample forecast statistics to compare with those for the models estimated using the standardised original data as shown in Table 2 in the body of the paper. Tables B2, below, provides the equivalent statistics for the models estimated from stationary versions of the causal variables.

B1. All models

The average error reduction across all models and estimation periods relative to the naïve models' forecasts was 1.5 percent based on *UMBRAE* (unweighted geometric mean), but an error increase relative to the naïve of 0.4 percent based on *CumRAE*. Those figures represent error *increases* compared to forecasts from models estimated using the original standardised series of 12.1 percent (*UMBRAE*) and 2.0 percent (*CumRAE*), or 10.8 percent based on individual model-by-estimation-period comparisons.

For the 36 model-by-estimation-periods shown in Table B2, average forecast errors were reduced relative to those from the models estimated using the original data for 11 models but *increased* for 25. Only one of the 11 error reductions was associated with the independent solar models' estimation period variations.

¹⁰ After transformation the three solar variables were random walks with drift, and the anthropogenic (CO₂ etc.) variable was trend stationary ($p=0.00$).

Table B2. NH temperature models estimated using data transformed to achieve stationarity: Fit, predictive validity, and bias

Model variables and estimation statistics				Forecasts: Number, Bias, Errors					Correlations			
	Putative causal variables†	1851/2 - (n)	\bar{R}^2	#	Bias‡	MdAE§	IQR§§	CumRAE*	UMBRAE*	Fit vs. Accuracy**	Bias vs. Error††	
All Land	Anthro	-	0.185		-0.70	0.51	0.69	1.080	0.983			
	Anthro	IPCC Solar	0.167	119	-0.70	0.52	0.71	1.081	0.985	-0.111	1.000	
	-	B2000 Solar	(49/48)		0.161	-0.64	0.49	0.62	0.968			0.925
	-	H1993 Solar			0.183	-0.63	0.49	0.65	0.962			0.919
	Anthro	-	0.165		-0.69	0.56	0.78	0.941	0.938			
	Anthro	IPCC Solar	0.168	69	-0.69	0.53	0.77	0.943	0.958	0.146	0.886	
	-	B2000 Solar	(99/98)		0.160	-0.72	0.62	0.87	0.993			0.971
	-	H1993 Solar			0.166	-0.73	0.61	0.87	1.005			0.992
	Anthro	-	0.084		-0.81	0.86	0.78	0.949	0.972			
	Anthro	IPCC Solar	0.076	49	-0.81	0.87	0.77	0.950	0.973	-0.054	0.917	
	-	B2000 Solar	(119/118)		0.082	-0.85	0.93	0.79	0.993			1.016
	-	H1993 Solar			0.080	-0.85	0.93	0.79	0.996			1.057
Anthro	-	0.145		-1.06	1.10	0.33	0.853	0.843				
Anthro	IPCC Solar	0.141	19	-1.06	1.12	0.32	0.857	0.846	0.991	1.000		
-	B2000 Solar	(149/148)		-0.006	-1.21	1.16	0.28	0.977			0.976	
-	H1993 Solar			0.005	-1.23	1.17	0.27	0.993			0.991	
Rural Land	Anthro	-	0.049		-0.63	0.48	0.98	1.588	1.431			
	Anthro	IPCC Solar	0.046	119	-0.64	0.57	1.10	1.691	1.480	-0.145	0.913	
	-	B2000 Solar	(49/48)		0.033	-0.44	0.39	0.62	1.106			1.068
	-	H1993 Solar			0.080	-0.44	0.42	0.71	1.172			1.250
	Anthro	-	0.105		-0.31	0.28	0.69	0.892	0.857			
	Anthro	IPCC Solar	0.097	69	-0.31	0.28	0.70	0.903	0.910	0.772	0.972	
	-	B2000 Solar	(99/98)		0.096	-0.39	0.35	0.67	1.075			1.062
	-	H1993 Solar			0.095	-0.40	0.34	0.68	1.091			1.154
	Anthro	-	0.093		-0.42	0.37	0.70	0.886	0.838			
	Anthro	IPCC Solar	0.087	49	-0.42	0.38	0.69	0.890	0.844	0.547	0.980	
	-	B2000 Solar	(119/118)		0.088	-0.50	0.52	0.69	1.044			1.082
	-	H1993 Solar			0.084	-0.50	0.51	0.66	1.018			1.030
Anthro	-	0.104		-0.75	0.80	0.51	0.845	0.786				
Anthro	IPCC Solar	0.098	19	-0.75	0.79	0.51	0.843	0.785	0.995	1.000		
-	B2000 Solar	(149/148)		0.050	-0.87	0.82	0.54	0.973			0.970	
-	H1993 Solar			0.049	-0.88	0.78	0.56	0.990			0.986	

† All models were estimated using ordinary least squares regression in STATA and include the variable "Volcanic".

‡ Mean signed error (forecast minus actual, °C).

§ §§ °C. **MdAE** is median absolute error. **IQR** is interquartile range and is calculated from signed errors.

* Cumulative Relative Absolute Error (Armstrong & Collopy 1992) and Unscaled Mean Bounded Absolute Error (Chen, Twycross, & Garibaldi 2017) figures are both relative to a naïve model forecast equal to the median value of the estimation data (a "no-change" forecast). Values of less than 1.0 represent error reductions relative to the naïve method (e.g., 0.95 represents an error reduction of 5%). Conversely values greater than 1.0 represent error increases (e.g., 1.20 represents an error increase of 20%).

** Sign-reversed Pearson correlation between the \bar{R}^2 s and the UMBRAEs.

†† Pearson correlation between the absolute values of Bias (°C) and the UMBRAEs.

B2. All Land models

The All NH Land temperature anomaly models all *increased* forecast errors on average compared to the errors of the out-of-sample forecasts from the original data estimated models. The error increases were 27.1 percent (AVL), 35.7 percent (AVSL), 65.6 percent (S_BVL), and 38.9 percent (S_HVL) based on geometric means of *UMBRAEs* across the four estimation periods in Table A2.

B3. Rural Land models (temperature data avoids heat islands)

The anthropogenic models (AVR and AVSR) forecasts from models estimated using stationary data were more accurate (*MdAE*, *CumRAE*, and *UMBRAE*) than the forecasts from the same models estimated using original data for estimation periods other than the longest, 1850 to 1999, period. For anthropogenic models estimated from the 1851/52 to 1899 stationary data, the increase in accuracy is more properly characterised as “inaccuracy reduction” relative to the naïve method.

Independent solar model (S_BVR and S_HVR) forecasts from models estimated using stationary data were uniformly *less* accurate (*MdAE*, *CumRAE*, and *UMBRAE*) than the forecasts from the same models estimated using original data and were *less* accurate than forecasts from the Naïve model for all but the models estimated from data to 1999.

B4. Summary and conclusions

Averaged across all four estimation periods and the two temperature target variables (NH All Land and NH Rural Land), the model which provided the most accurate temperature forecasts—i.e., reduced errors to the greatest extent relative to the naïve method) was the independent B2000 solar model (S_BVL and S_BVR) *when estimated from the original standardised data*. The error reduction (geometric mean of *UMBRAEs*) was 39.2 percent. The H1993 independent solar model (S_HVL and S_HVR), again when estimated from the original standardised data, was second best with an error reduction of 29.2 percent.

The anthropogenic no-solar model (AVL and AVR) *estimated from stationary data* come in a distant third with 6.0 percent error reduction. The anthropogenic model with the IPCC preferred solar variable (AVSL and AVSR) when estimated from stationary data came fourth with 4.5 percent error reduction.

When estimated from stationary data the B2000 and H1993 solar models provided forecasts that increased errors relative to the naïve method by 0.7 percent and 4.3 percent on average respectively. When estimated from the original standardised data the anthropogenic models provided forecasts that increased errors by 19.1 percent and by 16.5 percent when the IPCC solar variable was included.

The transformation of causal variables to achieve stationarity *increased* forecast errors on average and reduced the accuracy of forecasts from the best model. The independent solar model that included the B2000 solar variable reduced error by 39.2 percent on average when estimated from the original data, whereas the model that provided the most accurate forecasts when estimated from stationary data—the anthropogenic model without a solar variable—reduced errors by only 6 percent relative to the naïve model.