When are purely predictive models best?

Robert Northcott

to appear in *Disputatio*

Abstract
Can purely predictive models be useful in investigating causal systems? I argue "yes". Moreover, in many cases not only are they useful, they are essential. The alternative is to stick to models or mechanisms drawn from well-understood theory. But a necessary condition for explanation is empirical success, and in many cases in social and field sciences such success can only be achieved by purely predictive models, not by ones drawn from theory. Alas, the attempt to use theory to achieve explanation or insight without empirical success therefore fails, leaving us with the worst of both worlds – neither prediction nor explanation. Best go with empirical success by any means necessary. I support these methodological claims via case studies of two impressive feats of predictive modelling: opinion polling of political elections, and weather forecasting.

Keywords
Prediction, explanation, weather, causation, idealization

1. Introduction

Many areas of science have prioritized the development of theory and mechanisms with the aim of using them to explain messy and hard-to-predict field phenomena. In this paper I criticize this widespread methodological approach, arguing that instead we should prioritize empirical success – even though it may be difficult, and even if it means foregoing explanation. In the best cases, we can achieve both empirical success and explanation. But in many other cases we cannot, and then it is better to aim for empirical success without explanation than vice versa. Indeed, we have no choice: the possibility of achieving explanation without empirical success is an illusion, and moreover a harmful one insofar as it diverts effort away from the attempt to achieve empirical success. In some of the unpromising cases it does turn out to be possible to achieve both empirical

success and explanation after all, at least to a degree – but only via the route of empirical success first.

In many cases the only route to empirical success is via purely predictive models, i.e. models that do not attempt to capture a situation's causal structure and thus that may not deliver (causal) explanations. In these cases, purely predictive models are best.

I support these claims with two case studies of notable predictive success, namely forecasting of political elections and of weather. They reinforce the methodological lesson that (at least sometimes) predictive success should be prioritized over theory development and that, contrary to hopes, theory development cannot provide a shortcut to explanation without successful prediction first. After going through the case studies (sections 3 to 5), I return to the issue of prediction versus explanation and also consider the possibility of explanation via historical rather than forward-looking empirical success (sections 6 to 8).

2. Prediction versus explanation

Begin with a familiar and simple example: a Newtonian model of a cannonball dropped from a tower. Using this model, we may predict the acceleration of the ball and, thus, the time it takes to reach the ground. These predictions turn out to be very accurate. The model is causal: in particular, it describes how the Earth's gravity causally influences the ball's motion. Thus it gives a causal explanation of the ball's trajectory. Thus also, it is generalizable to many new cases – such as if the ball were a little heavier, or dropped from a different tower, or dropped at night rather than during the day.

Such a model is a paradigm case of idealization: it isolates and distorts particular causal factors and then uses its analysis of those factors to explain a real-world target, even though that target may contain many other factors that the model ignores. Here, the Newtonian model takes the ball and the Earth to be point masses, air resistance to be negligible, and so on. Taken literally, it is therefore false. Yet, as many philosophers of science have argued, idealized and therefore false models can nevertheless be explanatory. We do not need to go into the details of why; roughly, they all come down to the same verdict: idealized models can be explanatory when their falsity does not matter, i.e. when the idealizations are true enough for, say, predictive accuracy.[1] In our Newtonian case, the key vindicating feature is the accuracy of the model's predictions – it is this that gives warrant to its causal representation and thus to its causal explanations. It is also what gives warrant to the model's generalization to new contexts, since the causal structure it identifies is presumed to extrapolate. Whether the model will remain predictively accurate in a new context is then a matter of whether that context features significant other, disturbing causes unrepresented by the model. But even if for this reason the model is no longer predictively accurate, nevertheless we may still have warrant to accept that it truly identifies *some* of the causes present – if we have reason to think that the modeled causes continue to operate even in the presence of the unmodeled ones.

But this is the easy case, philosophically speaking: a causal model with empirical success, and which therefore offers both explanation and generalization. It is true that there are many such easy cases. But it is also true that there are many difficult ones. In particular, predictive success is often elusive. Indeed, this is arguably *typically* so in

[1] See, for instance, work by Nancy Cartwright, Daniel Hausman, Uskali Mäki, Michael Strevens, and Michael Weisberg. For an overview, see Weisberg 2013.

social sciences and in field sciences more generally.[2] What should we do then? This is where there is a great methodological split.

One common response, from philosophers and scientists alike, is that, roughly speaking, we should give priority to explanation over prediction. A long tradition, for instance, has doubted that systematic predictive success (or empirical success generally) in social science is possible. Among the reasons offered why not are: that social systems are *open*, i.e. are chronically subject to significant influences from non-social factors that are inevitably unmodeled by social science; that social systems exhibit *reflexivity*, i.e. that models themselves may influence their subject matter, thus creating a moving target; or simply that typically there are too many variables needing to be modeled (Taylor 1971, Giddens 1976, Hacking 1995, Lawson 1997). This pessimism has been bolstered by the great difficulty in practice of achieving predictive success. Thus, for instance, it has proved notoriously difficult to forecast what GDP or the unemployment rate will be in 12 months, who will win an election in six months, or (to cite a field but not social science) what the weather will be in one month.

In the face of this apparent impossibility, or at least great difficulty, of prediction, it has been argued that the goal of science should change: instead of prediction, it should instead be explanation. And the route to such explanation is the development of theory. Even in the absence of predictive success, it is held, such an approach offers the promise of understanding and insight. Indeed, such an achievement is often thought to be *superior* to mere predictive success, even when the latter is possible. The reason is that predictive success (or empirical success generally) in any particular case may require account to be taken of all factors, no matter how transient or sui generis. But what is of greater interest

[2] By 'field sciences' I mean non-laboratory investigations of systems that are not engineered artefacts.

is those factors or theoretical structures that generalize. It is the task of science, as a pursuit of systematic knowledge, to discover and isolate the latter. In the case of economics, for instance, we might be interested in a particular structure of choices and incentives, such as the Prisoner's Dilemma, because we understand this structure, and the outcomes to be expected from it, in terms of the discipline's fundamental building block of constrained agent rational choice, and because we think this structure crops up in many different places. Therefore, rather than get lost in local details it is more fruitful to focus on the Prisoner's Dilemma structure – *even though* it may predict accurately in hardly any particular cases. The analogy is with the methodological role of *mechanisms* in other sciences, such as neuroscience. Explanation in neuroscience, many philosophers have persuasively argued, is via appeal to mechanisms that, analogously to the Prisoner's Dilemma, are well understood and generalizable (Machamer et al 2000 and many others). Accordingly, just as in neuroscience, it is knowledge of mechanisms that enables true explanation and understanding in social science and field sciences generally, and our efforts should be allocated accordingly. It is knowledge of mechanisms, and the theory that underwrites it, that provides the understanding *of* any empirical success that we might achieve – and also provides understanding even in the many cases when our empirical success is imperfect.[3] Or so the argument goes.

Accordingly, there has been much philosophical support for the view that development of theory and mechanisms, rather than empirical or predictive accuracy, should be the primary goal of social science (Lawson 1997, Brante 2001, Elster 1989, Little 1991).[4] More than that, implicitly there has been much support too for this view from scientists themselves. The practice of economics, for instance, has been to develop a suite of formal

[3] There is a literature on whether "understanding", "insight" and the like have any epistemic value over and above explanation, but I will not address that here. I also do not commit myself to any particular definition of "mechanism".

[4] 'The proper function of a social science … is not prediction but diagnosis.' (Runciman 1963: 17).

models and then to apply these models as best we can to particular real-world targets, even knowing that rarely will they be fully accurate predictively. Usually, at least implicitly, such models are given a causal interpretation.[5]

So that is one response to the difficulty of prediction. Now turn to a different response – one that rejects a methodological emphasis on mechanisms and underlying theory (Cartwright 2007, Reiss 2008). One version of this response, motivated in part by case studies of empirical successes, has emphasized instead context-specific and extra-theoretical work, arguing that theories and mechanisms play at most a heuristic role (Alexandrova 2008, Alexandrova and Northcott 2009).

In this paper, I will investigate a specific variant of this alternative to the mechanist program. It advocates, roughly speaking, prioritizing prediction over explanation.[6] According to it, we should concentrate on achieving predictive success even if that means abandoning or moving beyond established theoretical models. This therefore offers the opposite methodological advice to that of the mechanists, and so in many cases implies a critique of scientific practice. The chief drawback of this alternative approach is that, because there may be no causal model underpinning a successful prediction, so we cannot offer any explanation of the predictive successes that we do achieve. Nor can we easily generalize from them in order to achieve predictive success in other cases too, including using them to assess hypothetical or counterfactual cases.

I will examine two examples of field phenomena that are highly complex but that have nevertheless been predicted successfully: political elections, and the weather. After

[5] Rodrik 2015 is a recent – and widely admired – articulation of this typical economist view.

[6] Of course, in practice prediction and explanation inform each other and are often closely related (Douglas 2009). Indeed, I will return to their relation later. Nevertheless, there remains a distinct methodological split.

looking at the methods used to achieve these successes, the lesson I will draw from them is that, contrary to the mechanistic view, the second methodological option above is preferable. That is, despite the difficulty of achieving accurate prediction, we should indeed prioritize it over explanation. Moreover, paradoxically, in the difficult cases this priority for prediction also turns out to be the only way to get explanations too. After that, I will then discuss the extent to which these conclusions generalize beyond the two examples.

3. Election prediction 1: polling aggregation

Successful election predictions are based on opinion polling, and the most successful predictors of all have been some *aggregators* of polling results. Take the 2012 US presidential election, in which Barack Obama defeated Mitt Romney. This campaign featured literally thousands of opinion polls. Famously, several aggregators correctly predicted the winner of all 51 states, as well as also getting Obama's national vote share correct to within a few tenths of a percent.[7] This is a stunning success, arguably with few equals in social science. Nor was it easy.[8] On the morning of the election Romney's odds at the bookmakers were 9/2, i.e. about 18%. Political futures markets such as InTrade put Romney's chances at about 28%. These market prices imply that common opinion was surprised by the outcome.[9]

---

[7] Four of the most successful aggregators were: http://votamatic.org, http://www.huffingtonpost.com/news/pollster/, http://election.princeton.edu/, and http://fivethirtyeight.blogs.nytimes.com/. The forecasting models for the first two of these were designed mainly by political science academics, the third by a neuroscience academic, and the last by a non-academic. Three of the four got every state right. Note: these predictions were from immediately beforehand; accuracy diminishes the further in advance of an election that predictions are made.

[8] See http://delong.typepad.com/sdj/2012/11/war-on-nate-silver-final-after-action-report-the-flag-of-science-flies-uncontested-over-silvergrad-weblogging.html for a list of 47 examples of failure, with an emphasis on their suspicion of polling-based prediction.

[9] There is an issue here about how we measure success in the case of probabilistic predictions. After all,

This success was not a fluke: the same poll aggregators have been successful in other elections too. And within any one election there have been many separate successful predictions, such as of individual Senate races or of margins of victory, which are at least partially independent of each other. Moreover, the aggregators' methods are independently persuasive.

The predictions here were the result of two distinct stages: first, opinion polling itself; then second, aggregation of individual polls. Begin with the former. In any opinion poll, the voting intentions of a sample serve as a proxy for those of a population. How might things go wrong, such that the sample will not be representative? The most well-known way is random sampling error: small samples can lead to misleading flukes. The larger the sample, the less this is a problem. But more important are various forms of sampling bias, i.e. of systematic sampling errors that (for a given sampling procedure) cannot be alleviated simply by increasing sample size.[10] Awareness of this crucial point lies at the heart of any serious election prediction. I'll illustrate here with an especially notable source of sampling bias, namely balancing for demographic factors.[11]

---

these market prices still made Obama favorite, so why should we term them 'surprised' by his victory? In reply, besides the simple fact of the overall winner, there were also relevant additional facts: who won each state; and how much they won them by. Odds-makers were not impressive with respect to these more detailed targets. Indeed, barring unlikely background assumptions, the details of the state-level results are hard to reconcile with a 28% chance of overall Romney victory. There is no serious dispute that the odds-makers and many other predictors were not accurate. On the further matter of which poll aggregator did best, see: http://rationality.org/2012/11/09/was-nate-silver-the-most-accurate-2012-election-pundit/.

[10] Lying in the background here are reference class issues. But given the unavoidable cognitive and epistemic constraints facing polling scientists, their choice of reference class is not arbitrary. And in practice the distinction between random and systematic sampling error is vital. Evidence of the latter's importance: almost 25% even of late polls miss the final election result by more than their official confidence interval, yet the expected miss rate given random sampling error alone should be only 5%. (http://fivethirtyeight.com/features/how-the-fivethirtyeight-senate-forecast-model-works/)

[11] Northcott 2015 briefly discusses some further sources.

Suppose, for example, that three-quarters of interviewees were women. Since there is good reason to think that women were disproportionately likely to vote for Obama, it follows that such a woman-heavy sample would give misleadingly pro-Obama predictions. This problem cannot be alleviated just by making the sample larger. Polling companies would therefore *rebalance* such a sample, in effect putting greater weight on men's responses. Such rebalancing is unavoidable if we wish to predict accurately, and every polling company performs some version of it. Any poll's headline figures are therefore heavily *constructed*. They are certainly not the raw survey results.

Exactly what rebalancing is required depends on assumptions about electoral turnout. For instance, in recent American presidential elections typically there have been slightly fewer men than women voters, so it would be a mistake to rebalance the sample to exactly 50-50. The exact figure may not be obvious, it needing to be inferred from imperfect data about past elections, and moreover with some assessment of how patterns of turnout might change again in the upcoming election.[12] Accordingly, different polling companies may reasonably choose slightly different rebalancing procedures. The result is the phenomenon of 'house effects', i.e. when a particular company's results systematically favor one or other candidate compared to the industry average. When assessing the significance of a poll for election prediction, it is vital to be aware of this.

The rebalancing issue is pressing because it applies to many other factors besides gender, such as: age; income; race; likeliness to vote; education; ownership of cellphones but not landlines; and home access to internet. Not only is the precise rebalancing procedure for each of these factors arguable, it is also arguable exactly which factors should be rebalanced for in the first place (Northcott 2015).

[12] Such rebalancing was behind the polling errors in both the 2015 and 2017 UK general elections.

Turn now to the aggregation of polls, which represents a second layer of method quite distinct from that required to conduct a single poll. Historically, poll aggregation has had a better predictive record than using individual polls alone. One obvious reason is that aggregation increases effective sample size and therefore reduces random sampling error. But it is not just that; it is also that sophisticated aggregation can mitigate the other sources of error too. This explains why the best aggregators beat simple averaging of the polls. Some polls are worthy of greater epistemic weight than others. Unless election day is very close, demographic factors can improve on the predictions of polls alone. Historical evidence can inform how much regression to the mean to expect if one candidate is unexpectedly far ahead, or if they have received a polling bounce from their party's convention. Some crucial issues during the 2012 campaign required such sophisticated analysis. Two prominent examples were: first, a persistent disagreement between state-level and national-level polling; and second, how likely it was the polls were wrong due to being systematically skewed against Romney. Neither of these issues could be addressed simply by taking a polling average, yet each was crucial to accurate prediction.[13] In sum, polling aggregation is a skillful matter, which is why only a few managed it successfully.

4. Election prediction 2: fundamentals modelling

There is an alternative approach to election prediction, and one that corresponds to the mechanistic strategy that prioritizes explanation over prediction. In particular, a literature in political science focuses on 'fundamentals', i.e. on variables that might be relevant to elections in general and not just to particular cases. These variables include economic

[13] See Northcott 2015 for detailed discussion.

conditions, the perceived extremism of candidates, incumbency, and so forth.[14] How does this literature fare?

Conveniently, it too has focused on US presidential elections. The sample size is relatively small, as fewer than 20 have good enough data. This creates a danger of overfitting. In response, models have typically featured only a small number of variables, most commonly economic ones such as growth in GDP, jobs or real incomes.[15] They are estimated on the basis of one part of the sample and then tested by tracking their predictive performance with respect to the rest of the sample. Even then, there remains a risk of overfitting: if a model predicts the first few out-of-sample elections quite well, will its success continue in future elections? Moreover, even if a model does successfully predict past elections, there is no guarantee the political environment is so stable that the model will remain correct in future too.

These caveats noted, it is true that the models can show some predictive success. On one estimate, the best ones' average error when predicting the incumbent party's share of the vote is between 2 and 3%.[16] But this is not quite as impressive as it might sound: first, for our purposes it is something of a cheat, in that one of the variables in by far the highest weighted model – Abramowitz 2008 – is a polling result, namely presidential approval rating. So the success is not achieved purely by fundamentals. Second, a 2-3% average error corresponds to an average error when estimating the *gap* between the

[14] Influential contributions include Fair 1978, Campbell and Wink 1990, Hibbs 2000, Abramowitz 2008, and Lewis-Beck and Tien 2008. Montgomery et al 2012 averages these and other models to achieve the best forecasting success of all.

[15] Literally thousands of economic variables could plausibly be deemed relevant, not to mention many non-economic ones too. The risk of overfitting is one good reason to prioritize prediction over retrospective accommodation. There is a long established literature in philosophy of science on the relative epistemic merits generally of prediction versus accommodation, but I will not discuss that here.

[16] See http://www.brendan-nyhan.com/blog/2011/11/a-comparison-of-presidential-forecasting-models.html for discussion and references.

leading two candidates of about 5%. And third, vote shares rarely deviate all that much from 50% anyway, so they are quite an easy target – indeed, another estimate is that economic variables account for only about 30-40% of the *variance* in incumbent party vote share.[17] Overall, the models do not predict individual election results very reliably. On many occasions they even get wrong the crude fact of which candidate won. For accurate prediction, it is necessary to incorporate the results of opinion polls.

Still, a primary motivation of the fundamentals literature, in keeping with the mechanistic strategy generally, is to be able to provide explanations. Can we establish *why* Obama won? But unfortunately the fundamentals approach fails on this score too.

On the polling side, in a trivial sense Obama's victory is 'explained' by the fact that, as revealed by aggregators, on the eve of the election a majority of the electorate were minded to vote for him. But, of course, for most investigative purposes a deeper explanation is required. Polling aggregation provides none.

On the fundamentals side, if its models had fared better they would have provided the very explanations that polling aggregation does not. After all, that is precisely the motivation for theory-centered methodology. Thus we might have been able to explain that Obama won because of, say, positive GDP and jobs statistics in the preceding two quarters. Unfortunately, though, the fundamentals models are not predictively accurate. And to explain requires identification of an event's causes, which requires a verified theory or causal model, which in turn requires empirical warrant.[18]

---

[17] http://fivethirtyeight.blogs.nytimes.com/2011/11/16/a-radical-centrist-view-on-election-forecasting/

[18] Following the literatures under consideration here, I focus on *causal* explanation. I do not mean to rule out the possibility of other forms of explanation.

Can the fundamentals models nevertheless provide us with explanations anyway? The argument would be that they have truly identified relevant causes. It might be postulated, for instance, that GDP or stock market growth does causally impact on voter preferences and thus on election outcomes. True, other causes impact too and so the models do not explain the outcomes fully nor predict them accurately, but that still leaves room for the claim that they explain them 'partially' by correctly identifying *some* of the causes present.[19]

But, alas, even this weaker claim is dubious here. First, the different models cite different variables. Abramowitz's, for instance, cites GDP growth, presidential approval rating, and a complex treatment of incumbency; Hibbs's though cites growth in real disposable income and the number of military fatalities abroad. Even among economic variables alone, some models cite GDP, some household incomes, some jobs data, some stock market performance, and so on, each with different combinations of lags. At heart, there are many different ways to achieve roughly the same limited predictive success, which shakes our faith that any one way has isolated the true causal drivers of election results. Perhaps the small sample size relative to the number of plausible variables makes this problem insoluble.

A second reason for pessimism is that, elsewhere in science, a standard response to such predictive failure is to test putative causes in isolation. As it were, at least we achieve predictive success in the isolated test. But unfortunately such experiments are impossible in the field case of election predictions. So as well as predictive failure at the level of elections as a whole, the causal factors picked out by the models have not earned their empirical keep by any other means either.

[19] See Northcott 2012 and Northcott 2013 for more on the relevant sense of partial explanation.

The upshot is that we have no warrant even for partial causes of election outcomes, and therefore no warrant even for partial explanations. Thus the basic conclusion stands: we have not achieved any explanation of election outcomes, and so the original motivation for turning to fundamentals models is frustrated.

5. Weather forecasting

On the face of it, the Earth's weather might seem an unlikely object of predictive success. It has long been thought a chaotic system, in other words that outcomes are indefinitely sensitive to exact initial conditions (Lorenz 1969). More recently, it has also been convincingly argued that weather predictions are (often) also indefinitely sensitive to *model* errors – that is, even tiny inaccuracies in a model can lead to very large errors in the predictions made by that model (Frigg et al 2014). These are obviously major challenges for prediction. Yet, despite them, weather forecasting accuracy has improved significantly over recent decades.[20] Hurricane paths are predicted more accurately and further ahead, and temperature and rainfall predictions are more accurate too. Overall, the reliability of seven-day forecasts now is equal to that of three-day forecasts 20 years ago (Bechtold et al 2012).[21] What explains this tremendous progress?

There are several factors. One is the huge improvement in the quality and quantity of weather data, stemming initially from the launch of the first weather satellites in the 1960s. There are now temperature, humidity and other reports of ever greater refinement both horizontally (currently increments of 20km squares) and vertically (currently 91

[20] Throughout, I will use the terms "prediction" and "forecast" interchangeably.
[21] I will concentrate on the work of the European Centre for Medium-Range Weather Forecasts (ECMWF), but similar remarks apply to other weather forecasters too.

separate altitude layers). Over 10 million observations per day are inputted into the ECMWF's calculations. A second source of progress has been the huge increase in available computing power. This has enabled ever more complex models to be used, ever more simulations to be run, and thus the huge improvements in the available data to be properly exploited. A third source of progress has been in the models of the weather themselves, which are used to make the predictions (see shortly). Finally, a fourth source has been improvement in analytical methods. Perhaps the most significant of these followed on from the introduction of stochastic terms in the basic model in the late 1990s: this allowed an *ensemble* method of forecasting to be adopted.[22] Multiple simulations are run – in the case of the ECMWF, currently about 50.[23] These are then used to generate probabilistic forecasts.[24]

The use of the ensemble method is particularly significant because it is the main way in which forecasters have overcome the problem of chaos. In particular, although any one forecast inevitably risks going seriously askew because of an arbitrarily small error in the inputted initial conditions, it has been found from experience that, as in many chaotic systems, these errors 'cancel out' over many iterations. That is, the errors are not systematically in one particular direction, and so the probabilistic forecasts derived from an ensemble of forecasts are not biased.

[22] Other analytical innovations include new forms of bias correction, which account for fluctuations in instrument calibration and enhance consistency between diverse types of observation – for instance, data from all of satellites, sea buoys, and conventional ground stations. Another important component is the balance struck between observation and background errors. These errors exhibit significant variations between observation types and locations. The balance struck between them determines the weight given to observations in the analysis (Bechtold et al 2012).

[23] The variety on which the ensemble works is generated not just by the model uncertainty introduced by stochastic terms in the model, but also by data uncertainty.

[24] A similar ensemble method is also used by leading polling aggregators.

These sources of progress have interacted with each other. For instance, the increase in data and computing power have enabled the development of more sophisticated models, the exploitation of which is constrained by the need to run the required number of simulations quickly enough to generate timely forecasts. Indeed the ensemble method of forecasting, notwithstanding its desirability, was simply not feasible until sufficient computing power became available. Meanwhile the needs of the model, and in particular knowledge from experience of what data improvements would most improve the accuracy of the model's predictions, in turn influence the gathering of data, such as the choice of instruments to be included on new satellites.

Turn now in more detail to the remaining source of progress, namely the improvements in the weather model itself.[25] At the heart of these models are differential equations that have been known for hundreds of years, namely Newton's laws of fluid dynamics. These are assumed to govern the fiendishly complex movements of air in the atmosphere, and how those are impacted by temperature, pressure, the Earth's rotation, the cycle of night and day, and so on. So far as is known, this 'fundamental' theory remains a true description of the weather system (or at least as approximately true as any other Newtonian model). However, in practice *it is not sufficient to generate accurate weather forecasts*. Moreover, refining the model from first principles has not proved to be an effective way to remedy this situation. Rather, a whole series of additions have had to be made to the model that reflect various sui generis factors. More to the point, the exact form that these additions should take is under-determined by fundamental theory. Instead, many different forms have been tried, and the ones adopted have been determined by a trial-and-error process. In particular, the huge amount of weather data now becomes an

[25] Much of the following is taken from the discussions of the ECMWF model in (Bechtold et al 2008) and (Jung et al 2010), as well as from personal communication with Roberto Buizza, Head of the Predictability Division at ECMWF.

important epistemic advantage because it is possible to test various tweaks on a vast archive of past data, as well as to test which tweak predicts best on new data. In this respect, the weather forecasters have it much easier than the election predictors, who were restricted to just 20 or so data points of past presidential elections.

There is a cost to this method though, familiar already from the elections case – although these changes to the model have greatly improved predictive accuracy, they have come at the cost of explanatory transparency. The reason is the usual one for purely predictive models, namely the deviation from well-established theoretical underpinnings. The various tweaks mean that there is no longer a warranted causal interpretation for each term in the model. Mathematically, the main differential equations now feature an extra term, not derived from theory, to represent new factors. The point is that the forms of these extra terms are not given by theory; rather, they are determined empirically, i.e. by whichever form gives the best predictive results. The same is true of deciding whether a particular factor should be incorporated into the model at all – sometimes, such as with ocean coupling (see shortly), theoretically well-motivated and initially promising innovations have eventually been dropped because they did not pay their way predictively.

In order to predict successfully, weather modelers must be aware of a huge range of detailed meteorological factors over and above the basic laws of fluid dynamics. Here are a couple of examples for illustration. First, consider *mountains*. These are well known to influence atmospheric circulation and to have large local effects on air flow and precipitation, both around mountains themselves and in surrounding areas. So it was realized in the early 2000s that introducing a term for the effect of mountains could potentially improve the model. The question was how exactly to do this. To work that out, it was necessary to move beyond theory:

> [One version of the model] included a 'cutoff' or 'effective' mountain height in the computation of gravity wave drag from the SSO scheme [i.e. the scheme to represent mountains]. The more physically realistic cutoff mountain height resulted in a decrease in gravity wave drag (GWD), reducing the excessive deceleration of flow over the Himalayas and Rocky Mountains ... However, climate runs showed an increase in the positive zonal wind bias over winter northern hemisphere mid-latitudes, suggesting that the reduction in GWD had been excessive. This problem was solved [in the next version of the model] by doubling the 'cutoff' mountain height and thereby increasing the amplitude of the gravity waves 'generated' by the SSO scheme by a factor of two. (Jung et al 2010: 9)[26]

Notice the sequence here: the *less* physically realistic formulation was the one eventually adopted, because it generated more accurate forecasts. That is, predictive fit trumped causal understanding. Instead, the role of causal understanding was a heuristic one – to *suggest* factors to be considered, in this case the impact on air flows of mountains. Exactly how those factors were best incorporated could not be derived from theory but instead was determined, in instrumentalist style, by brute predictive efficacy.

It is a similar story when considering whether a factor should be incorporated at all. Again, theory suggests – but prediction decides. Consider 'ocean coupling'. This refers, roughly speaking, to the way in which the ocean and atmosphere work together to transfer heat from the tropics to polar regions and also to influence the circulation of fresh and salt water. Key elements are the exchange of heat between sea and atmosphere, and the impact of ocean currents. Naturally, weather modelers sought to incorporate these apparently significant processes. But it was found that 'results from climate simulations

[26] The page references here and elsewhere in this section are to the versions of the papers published as ECMWF Technical Reports.

[that incorporated ocean coupling] are similar but slightly "worse" with respect to observations compared to the uncoupled simulations. Therefore, only results from the uncoupled simulations are shown' (Bechtold et al 2012: 22).[27] That is, a theoretical improvement was rejected on empirical grounds. This is the opposite of the mechanists' methodological recommendation; priority was given instead to improving predictive accuracy.

This pattern is general. For instance, another factor suggested by theory, which is being investigated at the moment, is the role of sea ice. But exactly how this should be incorporated into the model, or whether it should be incorporated at all, will be determined entirely by empirical considerations. Similar remarks apply to many other features, such as vertical diffusion, soil hydrology, clouds, vegetation, and ocean waves.

So the notable progress in weather forecasting over recent decades has not come from any improvement in fundamental theory, i.e. Newton's fluid dynamics, which indeed remains unchanged. Rather, it has been driven by relentlessly prioritizing predictive success even at the expense of explanatory transparency. The commercial imperative to generate accurate forecasts above all else, has focused minds methodologically.

Return now to the issue of model error. Recall: even minuscule errors in a model can lead to quickly escalating errors in that model's predictions. So how can this danger be averted, given that weather forecasters' current models are clearly not literally true and therefore are in error in the relevant sense? A purely theoretical solution is unavailable. The answer instead is a further advertisement for brute emphasis on predictive accuracy above all else. Simply put, many different versions of a model, including different

[27] Weather forecasters often refer to their models as 'climate' simulations. Nevertheless their focus is strictly on predicting weather, not on the long-term processes that are the focus of climate science.

stochastic adjustments, are tested against the empirical data. The ones selected are those that, as a matter of fact, predict best. This method has proven effective at avoiding the problem of model error – simply pick those models whose errors of representation turn out not to lead to errors of prediction.

An important methodological feature in weather forecasting is the *holistic* nature of its empirical testing. Particular tweaks are made to a model, and these are then tested by checking the model's *overall* predictive success. The reason is the presumed ubiquity of interactive effects: a tweak to the model might have one effect now but then a very different effect (or no effect) once other parts of the model are altered. Experience has suggested that such instability of effect is common. 'It is very difficult to understand how exactly changes in model formulation affect the climate of the model' (Jung et al 2010: 13). As a result, in practice it is impossible reliably to predict what the impact will be of a particular tweak. Rather, only testing can reveal the answer in any particular case.

This militates against causal inference, and so it is much more difficult for the weather forecasting models to give explanations than it is for them to give predictions. In this respect, their situation resembles that of the polling aggregators. (It also resembles that of the designers of some economic auctions – Alexandrova and Northcott 2009.) A (causal) explanation requires the claim that things would have been different if a particular factor had been different; but holistic prediction cannot license such detailed counterfactual claims.

In the weather case though, more so than in the election or auction cases, it seems there might be some partial exceptions to this 'no explanations' conclusion. At root, what makes these possible is the exceptional quantity of data available. I mention only one of these exceptions briefly here (taken from Jung et al 2010; see that paper and also

Bechtold et al 2008 for more details and examples): there have been extensive recent changes to the model's treatment of convection schemes in the tropics and to its treatment of the radiative properties of ice clouds. These changes have of course been thoroughly tested for their impact on predictive success and refined accordingly. But in addition, the data allowed modelers to test whether the two changes composed non-linearly or not. In this case, it was found that the non-linear – i.e. interactive – effects were relatively small. Accordingly, particular improvements in overall predictions now could justifiably be attributed to particular changes to the model. (I return to this issue in section 8 below.)

6. Prediction versus explanation revisited

Consider the following schematic table:

	Explanation	**No explanation**
Successful prediction	Slot 1: Newtonian cannonball	Slot 2: polling aggregation, weather forecasting
Unsuccessful prediction	Slot 3: (Empty)	Slot 4: fundamentals election prediction, much actual social science?

In Slot 1 in the table are the happy cases where we achieve the best of both worlds, such as when a causal model predicts accurately and thereby also causally explains. Successful cases of idealization, such as the Newtonian cannonball, fall into this category.

In Slot 2 are cases where we achieve predictive success, but only purely predictive success – i.e. when the methods and models required to achieve predictive success are

such that they do not yield generalizable explanations. Examples discussed in this paper are polling aggregation and weather forecasting.

In Slot 4 are cases where we get the worst of both worlds, i.e. no prediction and no explanation either. The fundamentals election models fall into this category: they offer causal models but their lack of empirical success means we have no warrant for them. Moreover, given the practical impossibility of testing the individual causes in isolation, it is not possible to warrant even the claim that they explain partially.

In Slot 3 are cases of explanation without prediction. None of our examples fits this description. I will argue now that that is no coincidence – this slot remains empty generally. The reason is that, according to all prevailing theories of explanation in philosophy of science, explanatory success requires empirical success. In particular, causal explanation requires true (or approximately true) identification of causes present. What can be the warrant for such identification? In science, it must be empirical success. The most convincing proof of such success in turn is successful prediction – or at least it is in cases where there exist many competing models and much scope for after-the-fact rationalization, which covers many cases of claimed "empirical success" in social science (Northcott and Alexandrova 2015) and arguably in field sciences generally.

Admittedly, there are two important caveats here. First, in other cases it may be that explanations can also gain warrant from empirical success achieved retrospectively (section 7). In those cases, Slot 3 is only empty if we understand "unsuccessful prediction" broadly as referring not only to forward-looking prediction but also to after-the-fact "retrodiction" too.

Second, care should be taken in interpreting "unsuccessful prediction" in another way as well. Sometimes a model may not predict accurately because of the influence of unmodeled disturbing causes, yet it may nevertheless still have truly identified some of the causes present, thus achieving partial explanation. But in such cases we still must have warrant for thinking that these causes are truly identified, and if that warrant doesn't come from successful prediction in the case at hand then it must come from somewhere else. In experimental sciences, controlled experiments elsewhere give us the needed warrant to believe that particular causes or mechanisms are present and behaving in the way our models suggest. But in field sciences such warrant is frequently absent.[28] That leaves us with no successful predictions of any sort, either of the situation at hand or of other relevant cases. This was precisely the problem with the fundamentals models of election prediction, and it leaves us stuck in Slot 4 with neither prediction nor explanation.[29] Thus, if "prediction" is interpreted broadly to mean "prediction *somewhere*" (i.e. to include relevant testing of proposed mechanisms by experiments in contexts other than the present one), then Slot 3 is necessarily empty. No empirical success, no explanation.

In this light, what can prioritizing mechanistic explanation over mere prediction (understood broadly as per the above) achieve? The answer is: a place in either Slot 1 or Slot 4, i.e. either both of prediction and explanation or neither of them. In the happy cases where predictive success is achieved, that means Slot 1, and all is fine. But when predictive success is not achieved, that means Slot 4. Yet Slot 4 is obviously inferior to Slot 2, which at least offers prediction even if still no explanation. The real problem is the

[28] To be sure, in recent decades there has been a growth industry of controlled experiments in economics and other fields. In principle, these might provide just the needed empirical warrant. In practice though, the problem of external validity has often hindered that (Northcott and Alexandrova 2015).

[29] Arguably, this problem is widespread. See Northcott and Alexandrova 2013 for an argument to this effect with respect to economics.

misguided belief that, by focusing on theory and mechanisms, we can at least ensure that we reach Slot 3 even if we don't achieve Slot 1, i.e. that we can at least achieve explanation even without empirical success. But this is to put the cart before the horse, and is how the idea that models can be explanatory without empirical success can become a harmful illusion. In a mistaken pursuit of Slot 3, we condemn ourselves to Slot 4 even though Slot 2 may still be achievable. That is, in fruitlessly pursuing explanation without prediction, we unnecessarily forego what may actually be achievable, namely prediction without explanation.

In sum, empirically successful models are always best, and whatever modelling strategy yields them is the one we have to follow. And when are *purely* predictive models best? Answer: when otherwise empirical success is unobtainable. Something is better than nothing.

7. Historical explanation[30]

If we understand prediction narrowly to refer only to claims about future events then explanation *is* possible without it – because, as mentioned, sometimes it can be achieved by historical analysis too. Even if I failed to predict beforehand that I would win my tennis match, say, still I might successfully be able to explain my win afterwards by noting that my opponent got injured halfway through.

In response, note again that empirical success is still necessary for explanation. In particular, in the case of causal explanation, true identification of a relevant cause is

[30] I thank the editors and an anonymous referee for pressing me on the issues in this section.

necessary – if my opponent had not got injured, and if that injury had not caused their defeat, my explanation would have been incorrect.

Whether the empirical success is via predictions before the fact or via historical analysis after it, the interesting issue, methodologically speaking, remains: how do we achieve such success? Return to our two case studies, beginning with elections. How might we explain election results after the fact? Not by appeal to the fundamentals models, for the reason pointed out earlier: there is no warrant for claiming they have truly identified even some of the causes of election outcomes. An alternative is after-the-fact questionnaires, when voters are asked why they voted as they did. Certainly, these are potentially good evidence for explanations of an election result, although they do come with a couple of caveats: first, people's answers to such questionnaires are often unreliable. (Indeed, questionnaire answers are often inconsistent with recorded vote totals.) Second, so far at least, explanations drawn from such questionnaires have lacked generalizability: no one has been able to construct from them a model with significant predictive power for elections in general, as opposed to constructing after-the-fact explanations, varying each time, for particular cases.

Another approach is to use demographics. In the context of American presidential elections, the ambition is to explain election outcomes in particular states (and nationally) by reference to those states' population profiles with respect to ethnicity, wealth, levels of education, and so on. As mentioned, such demographic variables do add predictive value several months ahead of time, although polls dominate them by the time of the election date itself. But can demographic models help with explanation? To a degree, yes – and that degree is precisely the degree to which their empirical success can be established. For example, comparing election results in different American states in effect offers the possibility of a natural experiment. With due care, it can therefore be established that,

say, the different percentage of black Democratic voters in Wisconsin and South Carolina explains a certain amount of the difference in Clinton's vote share in those two states during her 2016 primary campaign against Sanders.[31]

In the weather forecasting example, by contrast, it is hard to see any generalizable explanations beyond those limited ones inferable from the forecasting models (section 8 below).

Where does this leave us? I think with two conclusions, both of which reiterate the paper's main claims. First, explanation does require empirical success, but this success can sometimes take the form of retrospective accommodations as well as future predictions. If "prediction" is interpreted broadly in this way, Slot 3 remains empty. The second conclusion is methodological: an emphasis on the development of theories and mechanisms is often a poor way to achieve this empirical success. In such cases, purely predictive models are best, although sometimes case-specific historical explanations might also be available. That said, in other cases theory may obtain sufficient empirical success to warrant explanations or at least partial explanations after all – one example is demographic analysis of elections. Still, even then they only do so thanks to empirical success, in other words Slot 3 again remains empty.

8. Coda: in further defense of prediction

Sometimes there may even be scope for moving from Slot 2 to Slot 1, i.e. for a purely predictive model also to generate explanations. At the end of section 6, I mentioned a

[31] Note that this would yield only a partial explanation of the primary results, since it identifies only one cause of them. Moreover the explanation is a little imprecise, both because the definition of a 'black' voter is imprecise and because the impact of ethnicity on voting patterns is not completely stable across states.

couple of examples of this in the weather forecasting case, or at least of tentative steps in that direction. If the quantity of data is high, then it is more possible to run natural experiments in order to test non-holistically the impact of particular parts of a model. In such cases we may indeed acquire warrant for certain causal – and thus explanatory – claims, and we would be moving into Slot 1.

Being in Slot 1, the predictive model would no longer be *purely* predictive. But the point is the route by which Slot 1 is reached. It is only by an initial methodological prioritization of prediction that sufficient empirical success is achieved to enable the subsequent generation also of explanations. To go for explanation straightaway, by sticking to a model with well-known mechanistic operation even in the face of predictive shortfall, would be exactly the wrong path, analogous to the case of the fundamentals models of elections. We need to be prepared to depart from established theory in the service of prediction as ultimately this turns out to be the only hope in these cases also of achieving explanation too.

Such a methodological approach reflects the heuristic view of models mentioned in section 2, according to which theory alone should *not* be seen as providing models that can be confirmed and generate explanations. Instead, the role of theory is merely to suggest inputs for the extra-theoretical – usually empirical and case-specific – work needed to generate eventual causal hypotheses. Accordingly, the initial theoretical models themselves are not explanatory; rather, what might be explanatory are the independent extra-theoretical "models" that eventually result (Alexandrova 2008, Alexandrova and Northcott 2009). Weather forecasting is a fine example of this.[32]

[32] Some, such as Kuorikoski and Ylikoski 2015, see even heuristic models as indirectly explanatory to a degree. Perhaps. But regardless of one's preferred semantics of "explanatory", the key point of difference here remains, since Kuorikoski and Ylikoski still support a methodological emphasis on explanation over

There are also two additional senses in which even purely predictive models can sometimes do more than merely predict. The first is that they often license successful *interventions* – and thus causal knowledge of a kind. One example of this is the many instances of chemotherapy where the mechanism is unknown but the efficacy is well established. Another example, already mentioned, is economic auctions, and in particular the well-studied case of government-run spectrum auctions (Guala 2005, Alexandrova 2008), where the success of the auction design is well established even though the holistic nature of the pre-testing required to generate that design means we have no detailed mechanistic explanation of the auctioneers' success.

True enough, in both the chemotherapy and auction cases extrapolation of the models to new applications is sadly difficult, requiring much fresh laborious work each time (Alexandrova and Northcott 2009). This is the disadvantage of being in Slot 2 rather than Slot 1: the predictive success is not easily generalizable. Nevertheless, especially in the auctions case, predictive success in one application has enabled success in some new applications to come at least a little more easily.

Indeed, a similar story applies in the election case too: the polling aggregators' success in one election did generalize to other elections – to some degree. Admittedly though, the case-specific nature of polling aggregation means that a serious aggregator must build a new election prediction model each time, and as a result, although sometimes the aggregators were again successful, a first success did not guarantee a second (Northcott 2015).

---

prediction.

Nevertheless, overall the point is that sometimes some of the benefits of Slot 1 can be achieved, albeit imperfectly and more laboriously, by going via Slot 2: in the case of weather forecasting some degree of explanation, and in the other cases some degree of generalizability. The common feature is that such benefits as do accrue rest crucially on predictive success. This remains the foundation for everything else. For this reason, reaching Slot 1 via Slot 2 remains superior to doomed attempts to progress via Slot 3. In other words, when our models are not predictively accurate, just refining those models according to theoretical criteria is an unpromising strategy. In these cases, progress can only be made by the hard empirical work of improving predictions instead. That is when purely predictive models are best.

References

Abramowitz, Alan. 2008. It's About Time: Forecasting the 2008 Presidential Election with the Time-for-Change Model. *International Journal of Forecasting* 24: 209-217.

Alexandrova, Anna. 2008. Making Models Count. *Philosophy of Science* 75: 383-404.

Alexandrova, Anna and Robert Northcott. 2009. Progress in economics. In *Oxford Handbook of Philosophy of Economics*. Oxford: Oxford University Press.

Bechtold, Peter, Martin Koehler, Thomas Jung, Francisco Doblas-Reyes, Martin Leutbecher, Mark Rodwell, Frederic Vitart, and Gianpaolo Balsamo. 2008. Advances in Simulating Atmospheric Variability with the ECMWF Model: From Synoptic to Decadal Time-scales. *Quarterly Journal of the Royal Meteorological Society* 134: 1337–1351. (ECMWF Technical Memorandum No 556)

Bechtold, Peter, Peter Bauer, Paul Berrisford, Jean Bidlot, Carla Cardinali, Thomas Haiden, Martin Janousek, Daniel Klocke, Linus Magnusson, Tony McNally, Fernando Prates, Mark Rodwell, Noureddine Semane, and Frederic Vitart. 2012. Progress in Predicting Tropical Systems: The Role of Convection. ECMWF Research Department Technical Memorandum no 686.

Brante, Thomas. 2001. Consequences of Realism for Sociological Theory-Building. *Journal for the Theory of Social Behaviour* 31: 167–194.

Campbell, James and Kenneth Wink. 1990. Trial-Heat Forecasts of the Presidential Vote. *American Politics Quarterly* 18: 251–269.

Cartwright, Nancy. 2007. *Hunting Causes and Using Them*. Cambridge: Cambridge University Press.

Douglas, Heather. 2009. Reintroducing Prediction to Explanation. *Philosophy of Science* 76: 444-463.

Elster, Jon. 1989. *Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press

Fair, Ray. 1978. The Effect of Economic Events on Votes for President. *Review of Economics and Statistics* 60: 159-173.

Frigg, Roman, Seamus Bradley, Hailiang Du, and Leonard Smith. 2014. Laplace's Demon and the Adventures of His Apprentices. *Philosophy of Science* 81: 31-59.

Giddens, Anthony. 1976. *New Rules of Sociological Method: A Positive Critique of interpretative Sociologies*. London: Hutchinson.

Guala, Francesco. 2005. *Methodology of Experimental Economics*. Cambridge: Cambridge University Press.

Hacking, Ian. 1995. The Looping Effect of Human Kinds. In *Causal Cognition an Interdisciplinary Approach*. Oxford: Oxford University Press.

Hibbs, Douglas. 2000. Bread and Peace Voting in US Presidential Elections. *Public Choice* 104: 149–180.

Jung, Thomas, Gianpaolo Balsamo, Peter Bechtold, Anton Beljaars, Martin Koehler, Martin Miller, Jean-Jacques Morcrette, Andrew Orr, Mark Rodwell, and Adrian Tompkins. 2010. The ECMWF Model Climate: Recent Progress Through Improved Physical Parametrizations. *Quarterly Journal of the Royal Meteorological Society* 136: 1145–1160. (ECMWF Technical Memorandum No 623)

Kuorikoski Jaakko, and Petri Ylikoski. 2015. External representations and scientific understanding. *Synthese* 192: 3817-3837.

Lawson, Tony. 1997. *Economics and Reality*. London: Routledge.

Lewis-Beck, Michael, and Charles Tien. 2008. The Job of President and the Jobs Model Forecast: Obama for '08? *PS: Political Science and Politics* 41: 687-690.

Little, Daniel. 1991. *Varieties of Social Explanation*. Boulder: Westview.

Lorenz, Edward. 1969. Three Approaches to Atmospheric Predictability. *Bulletin of the American Meteorological Society* 50: 345–349.

Machamer, Peter, Lindley Darden, and Carl Craver. 2000. Thinking About Mechanisms. *Philosophy of Science* 67: 1-25.

Montgomery, Jacob, Florian Hollenbach and Michael Ward. 2012. Ensemble Predictions of the 2012 US Presidential Election. *PS: Political Science and Politics* 45: 651-654.

Northcott, Robert. 2012. Partial Explanations in Social Science. In *Oxford Handbook of Philosophy of Social Science*. Oxford: Oxford University Press.

Northcott, Robert. 2013. Degree of Explanation. *Synthese* 190: 3087-3105.

Northcott, Robert. 2015. Opinion Polling and Election Predictions. *Philosophy of Science* 82: 1260-1271.

Northcott, Robert and Anna Alexandrova. 2013. It's Just a Feeling: Why Economic Models do not Explain. *Journal of Economic Methodology* 20: 262-267.

Northcott, Robert and Anna Alexandrova. 2015. Prisoner's Dilemma Doesn't Explain Much. In *The Prisoner's Dilemma*. Cambridge: Cambridge University Press.

Reiss, Julian. 2008. *Error in Economics: Towards a More Evidence-Based Methodology*. London: Routledge.

Rodrik, Dani. 2015. *Economics Rules: the rights and wrongs of the dismal science*. New

York: Oxford University Press.

Runciman, W.G. 1963. *Social Science and Political Theory*. Cambridge: Cambridge University Press.

Taylor, Charles. 1971. Interpretation and the Sciences of Man. *Review of Metaphysics* 25: 3-51.

Weisberg, Michael. 2013. *Simulation and Similarity: Using Models to Understand the World*. New York: Oxford University Press.

All cited 2014 URLs were accessed in November 2014, all earlier ones in August 2013.

Acknowledgements

I am grateful for helpful feedback from two anonymous referees and from Maria Jimenez Buedo and Federica Russo.