Electronic Supplementary Material (ESI) for Environmental Science: Atmospheres.

Measuring the Impact of Air Quality Related Interventions

Karl Ropkins [1*], James E. Tate [1], Anthony Walker [2], Tony Clark [2]

[*] Corresponding Author Karl Ropkins, email: k.ropkins@its.leeds.ac.uk; [1] Institute for Transport Studies, University of Leeds, Leeds, LS2 9JT, UK; [2] Joint Air Quality Unit, Department for Transport & Department for Environment, Food and Rural Affairs, Marsham Street, London, SW1P 4DF, UK.

Supporting Information

Section 1: Additional Site Information

The Headingley air quality monitoring station is Automatic Urban and Rural Network (AURN) affiliated, and further information about the site can be found at:

https://uk-air.defra.gov.uk/networks/site-info?site_id=LED6

Other local air quality monitoring stations used in this study (Kirkstall Road and Temple Newsam) are independently operated by Leeds City Council (LCC), and further information about these, other air quality monitoring and management activities undertaken by LCC can be found in, e.g.:

https://cleanairleeds.co.uk/sites/default/files/Leeds%20ASR%202018.pdf

Figure S1 shows the locations of all eleven automatic air quality monitoring stations in the Leeds area.

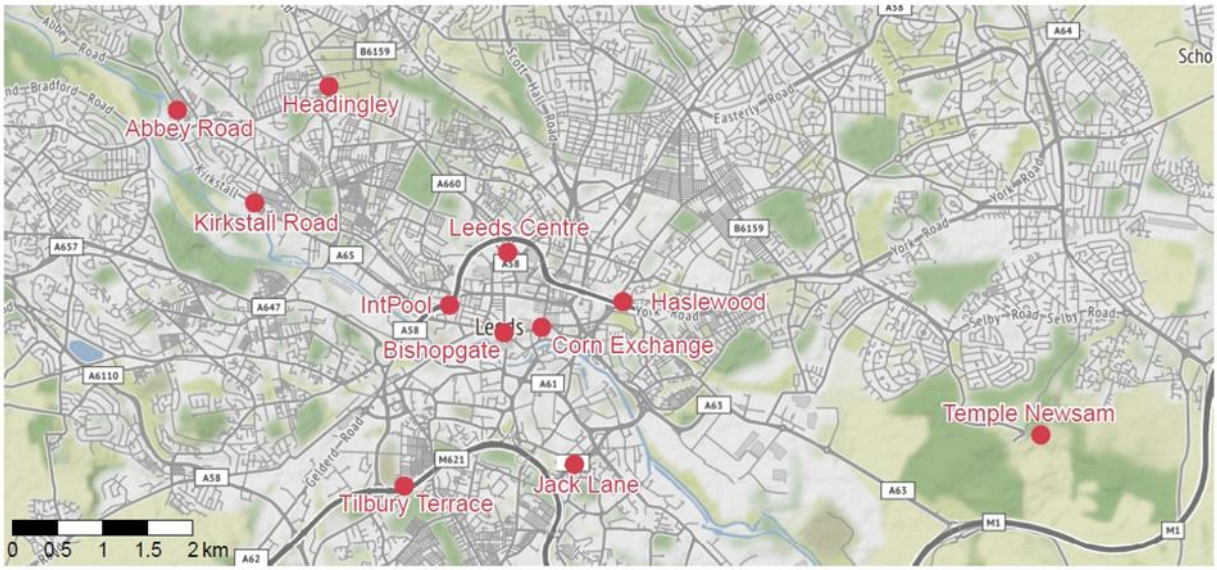


Figure S1: Locations of Leeds City Council (LCC) and Automatic Urban and Rural Network (AURN) monitoring stations in Leeds. Leeds Centre and Headingley are AURN and AURN affiliated sites, respectively, and all other sites are LCC operated sites. Headingley (the intervention site), Kirkstall (used as nearby control) and Temple Newsam (the local background) were used in the main study, but analyses of other sites are included below as part of further discussion of events. (Map tiles produced by Stamen Design, under CC BY 3.0. Data under ODbL using R package OpenStreetMap; Fellows & Stotz[50].)

Section 2: Theil-Sen Analysis of NO_2 Trends at Main Study Sites

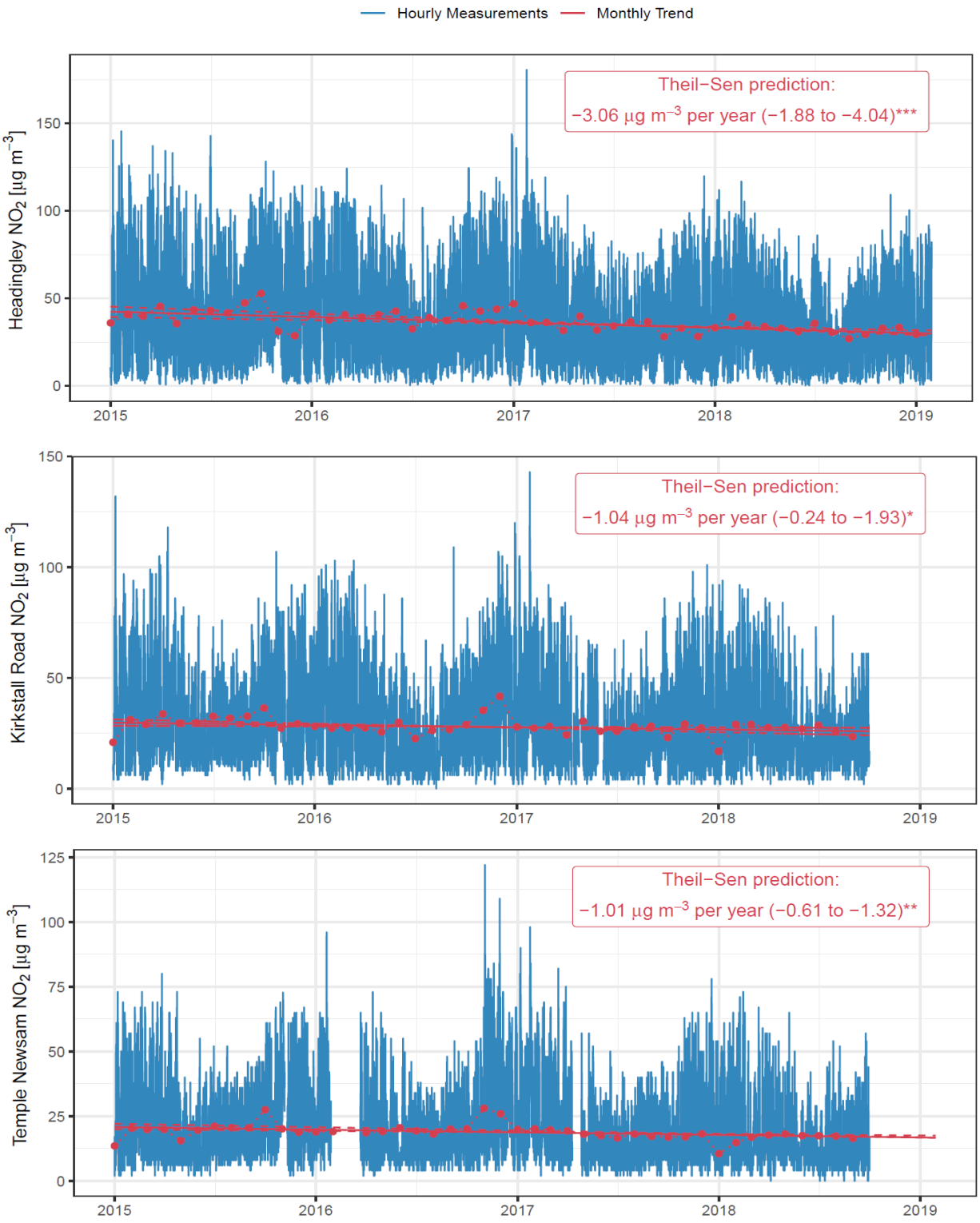


Figure S2: Ambient NO_2 1-hour resolution time-series for Headingley, Kirkstall Road and Temple Newsam (blue), with estimated underlying general trend (red) estimated using deseasonalised month-average measurement and Theil-Sen method in openair[43] [*** p<0.001, ** p<0.01, * p<0.05].

Section 3: Selection of Main Study Sites

Although all eleven sites were investigated as part of this study, work reported here focuses on these three sites because the objective here is to demonstrate the measurement of the impact of a local intervention. The local impact measurement itself is made using Headingley, meteorological and background data. Headingley is on the outskirts of Leeds. By comparison to Kirkstall, other sites were significantly less like 'Headingley-without-the-intervention'. For example, several were inner-city sites and/or in residential areas where bus contributions from multiple near-by roads/routes were complex, several had less complete time-series than Kirkstall, and all were further away from Headingley than Kirkstall. So, Kirkstall was selected as 'best available control'.

Section 4: Break-point/segment Analysis of NO_2 at Other Sites

The analysis of data from other near-by sites is reported here in Figures S3 and S4 and briefly discussed as part of the interpretation of the likely nature of the earlier (2015) break-point/segment seen in the Headingley data.

A 2015 break-point/segment, similar in both timing and magnitude to that at Headingley, was also observed at Tillbury Terrace and Intpool. At several other sites, e.g. Jack Lane and Corn Exchange, a 2015 break-point was detected but dismissed by either BIC or p-score screening (see Methods, section 2.2). No 2015 events were seen at other sites. However, in all cases, except Leeds Centre and Temple Newsam, there was significant missing data late 2015 and/or 2016, which may hinder analysis in break-point/segment detection and quantification in this period. In the cases of two sites, Abbey Road and Bishopgate, at-site monitoring only started late 2016 and early 2017, respectively, so no comment can be made regarding the likelihood of a 2015 event at these. As late 2015 events are seen at several sites, and, where seen, they were of magnitudes close to the estimated detection limit for the method, it suggests it is a less consistently/confidently detected but potentially more widespread than the bus intervention seen in 2018. Although far from unambiguous, it not being observed at the background site, Temple Newsam, suggests it may be urban rather than background/regional in nature. This is about the time Euro 6 vehicle regulations were introduced in the UK and others[e.g.,32] have reported similar changes that aligned with the introduction of earlier vehicle regulations. However, at this stage without further work and the analysis of more sites across the UK, such an interpretation would be highly speculative.

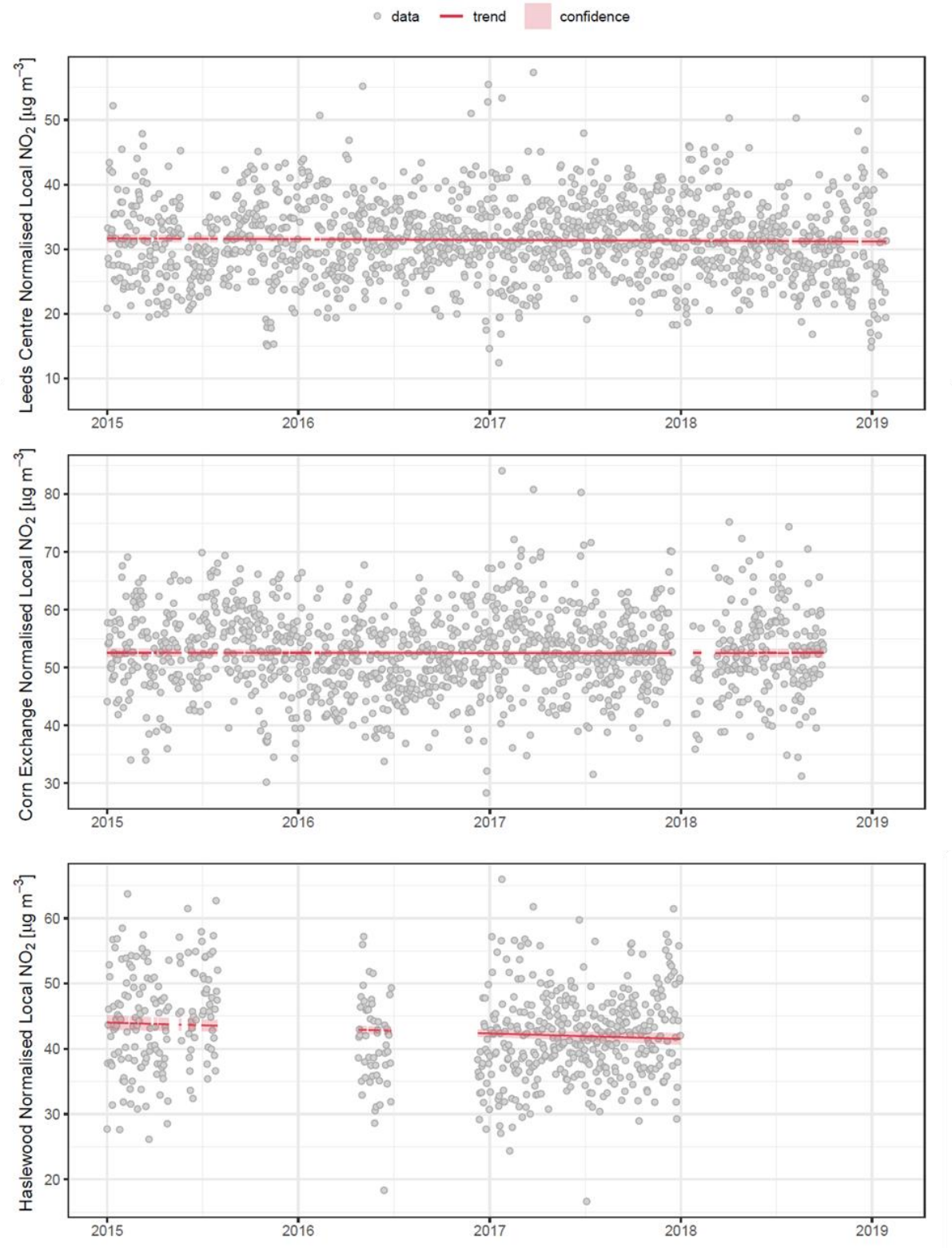


Figure S3: Break-point detection and change-segment analysis of normalised local contributions for Leed Centre, Corn Exchange and Haslewood time-series (Top, Middle and Bottom); data (grey), change-segments, with start and ends marked (blue) and associated confidence intervals (blue dashed) and segmented trends (red).

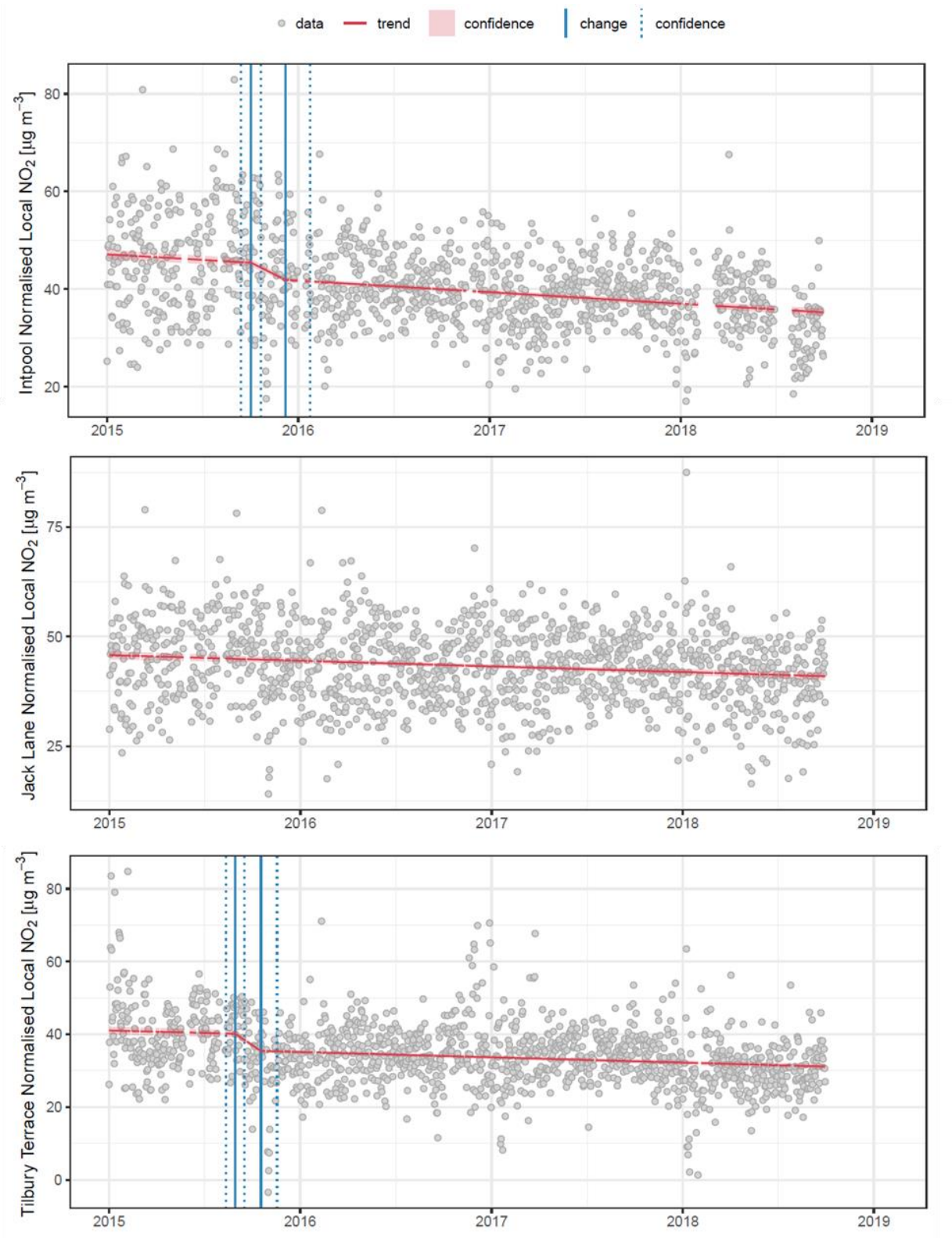


Figure S4: Break-point detection and change-segment analysis of normalised local contributions for Intpool, Jack Lane and Tillbury Terrace time-series (Top, Middle and Bottom); data (grey), change-segments, with start and ends marked (blue) and associated confidence intervals (blue dashed) and segmented trends (red).

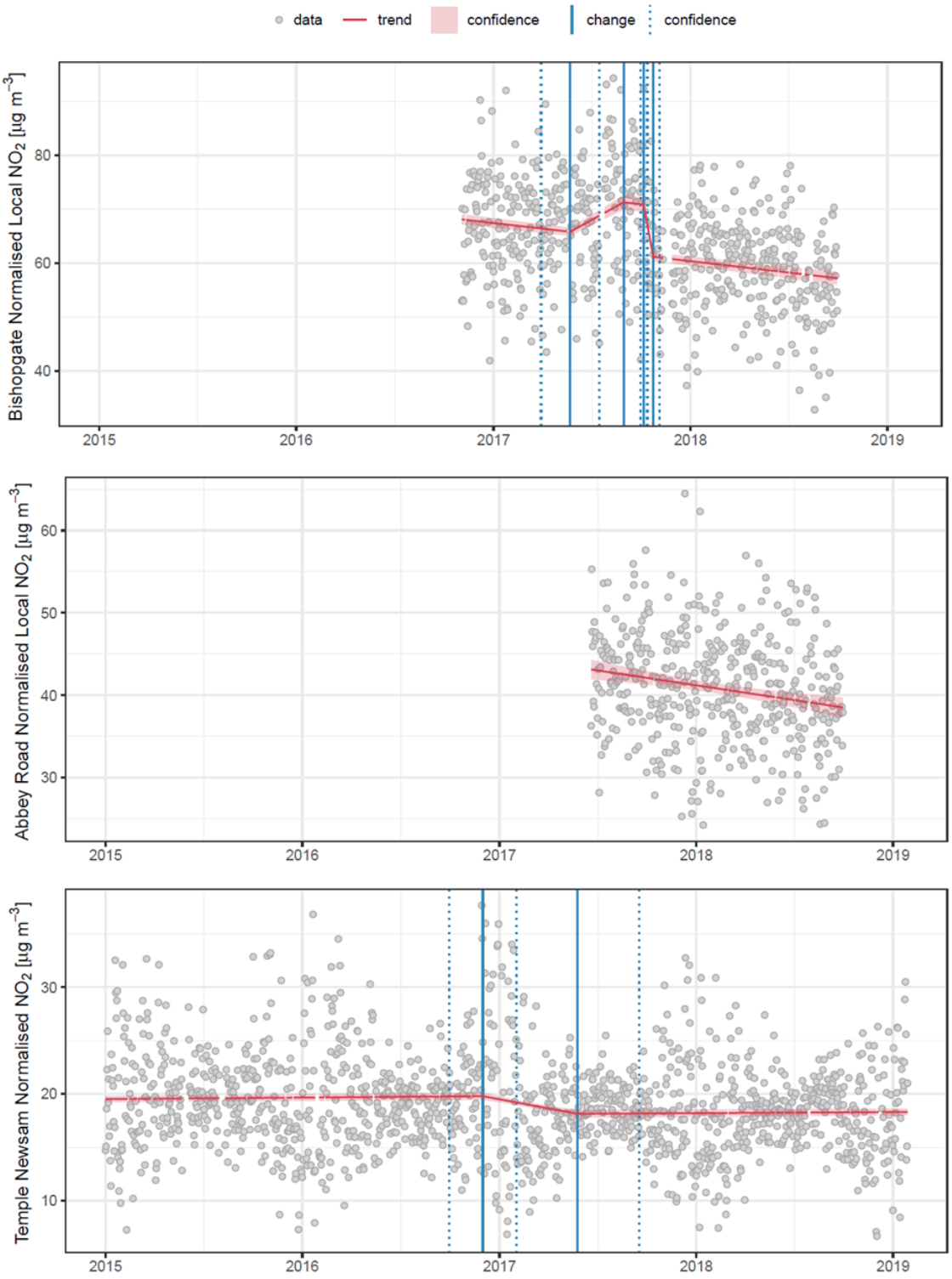


Figure S5: Break-point detection and change-segment analysis of normalised local contributions for Bishopgate, Abbey Road and Temple Newsam time-series (Top, Middle and Bottom); data (grey), change-segments, with start and ends marked (blue) and associated confidence intervals (blue dashed) and segmented trends (red). Temple Newsam analysis is of the modelled data used as the background for all other sites and therefore itself without background subtraction. Given the limited amount of data available for Bishopgate, the 2017 event detected there should be treated with caution.

Section 5: Simulation Study Characterisation of Break-point/segment Methods

Traditionally, change detection methods have been applied to air quality applications in relative isolation: methods applied, results reported and interpreted on the basis of what was expected (e.g. seen elsewhere or predicted using modelling). This is entirely understandable because for the most part, the methods are applied to applications where there is little or no other evidence regarding performance. However, there is a need to extend research efforts and investigate the likely performance of change-detection methods if we are to ask the authorities tasked with the delivery of air quality improvements to use these methods to benchmark their air quality management activities. There is also a need to consider the trade-offs between minimal and more-aggressive signal isolation when used as a prelude to break-point detection because: (1) Increasing method complexity creates extra burdens for users, e.g. more datasets to collect and quality assure, and more sophisticated models to implement. And, (2) These steps by their nature remove variance, and variance is the data property that change-detection methods typically test for change. When applied in a standalone fashion, more sophisticated, multiple-site methods that filter data and smooth trends over longer time-periods are obviously highly useful when investigating underlying regional and/or longer time-scale processes. But these methods can also potentially distort outcomes when used to pre-process data prior to an analysis like break-point testing if, for example, models are over-fit and variance associated with local change is removed. With this in mind, data sources, data handling, method refinements and simulation testing strategies are all discussed as part of this work to provide measures of both intervention impact and method performance, and to contribute to efforts to develop more widely applicable environmental change-point detection and quantification methods in future.

The findings reported (Section 3.2 in the main paper) are in good agreement with expected findings for a Euro VI bus fleet upgrade, so appear sensible. This is encouraging, but if such methods are to be used more widely benchmark the performance and appraisal of air quality interventions, the methods need to be verified and characterised.

A series of simulation studies were undertaken to investigate the performance of the methods used. Elsewhere[e.g.*] it has been observed that even the best designed simulations are artificial constructs and that analytics tend to work better on simulated data than on real-world data. With this in mind, a study-specific approach was adopted for this work. For the first set of simulations, the Headingley time-series was used to build the base case for simulation: The above analysis was repeated, and detected break-points were subtracted from the local contribution time-series and the time-series rebuilt. This generated a base case which when analysed contained no detectable break-points, but had statistical properties (mean, variance, etc.) that were

* Additional reference (not in main paper): B. M. Kim and R. C. Henry, Extension of self-modeling curve resolution to mixtures of more than three components: Part 3. Atmospheric aerosol data simulation studies, Chemom. Intell. Lab. Syst., 2000, 52, 2, 145–154, DOI: 10.1016/S0169-7439(00)00077-0.

highly similar to the original Headingley time-series. A change was simulated by isolating the local contribution of this base case, adding an artificial change at a known point and of a known duration and magnitude, rebuilding the time-series, and then rerunning the analysis and comparing the change and change prediction (Figure S6). For the first step of the simulation exercise, 2000 randomly selected changes were simulated using the rules: change anywhere in data range, magnitude and duration randomly selected from the ranges +50% to -50% and 1 to 100 days, respectively.

Figure S7 summarises the outcomes for simulated instantaneous changes (classic break-points) and Figures S9 and S10 summarise outcomes for simulated changes of variable durations. Unsurprisingly, better performance is observed for instantaneous change detection by comparison to more gradual change (compare S7 and S9 or see Figure S8). However, Figure S9 Top Right demonstrates the performance of the method when detecting the location of simulated change-segment mid-point. The high agreement (nearest to y = x, predicted = actual) for the majority of cases, demonstrates that the approach is actual still very good at locating the point about which change happens even when that change is more gradual. Horizontal regions at the start and end of the plot range reflect failure to detect changes that happened within the range of the first or last averaging window (10% of the time-series range) used when break-point testing. Other deviations from y = x typically associate with smaller magnitude changes, roughly -10% to +10%, and reflect increasing misassignment as the method detection limit is approached (see below). Figure S9 Top Left and Bottom Left show similar plots for change-segment start and end prediction, respectively. Although agreement is not as good here, this is to be expected because identifying the start or end of a gradual change is obviously harder than detecting either instantaneous change or gradual change mid-point. One feature worth noting here is that the method tends to assign starts late and ends early and that this trend becomes more pronounced as change magnitudes get smaller. Figure S9 Bottom Right shows magnitude predictions. Here, again, agreement is good in the majority of cases, and cases that deviate from y = x tend to associate with the start and end of the time-series, indicating a change happen within or near to the first or last averaging windows. One other feature to note here is the gap in the middle of the data range which is associated with undetected changes. This indicates an asymmetrical detection limit of *ca.* -5% for decreases and *ca.* 10% for increases at Headingley during the timescales of this study.

This simulation approach can also be used to investigate more complex situations. For example, when testing multiple break-points scenarios, it was noted that near predictions were often influenced by the break-point test window size. As the distance between two break-points decreases to window length, one, often the smaller, or both would be displaced, and as distance between break-points decreased to less than the window length, the two were often merged and detected at a single mid-location change or the smaller was sometimes obscured. There is also some indication that (in simulation at least) wrong break-point assignments are often less stable that than correct assignments. So, break-point consistency when an analysis is repeated with a

shorter or longer time-series, and/or different time windows, may also provide an extra measure of likely performance.

The effect of the properties of the supplied data on method performance was also investigated by modifying the properties of the base case and repeating this simulation procedure multiple times (typically 10 per investigated parameter). To compare the changes in method performance across a series of these simulation sets (each equivalent to Figure S9), a simplified descriptor was applied: % near fit, which was the percentage of predictions with ± 2 months of actual date for start and end points and within ± 10% of prediction for magnitude (Figure S11). Selected outputs from this simulation study are shown in Figures S12-S14. Here, general observations are consistent with those reported above. Predictive power generally decreases significantly towards the start and end of the study period (indicated by red low near fit regions in Figure S12 Left Top and Left Bottom plots and in other examples in the supporting information), reflecting the expected lower performance in the ranges of first and last break-point test window. Here, there is also some variation in performance with data properties such as gradient and variance. However, away from the time-series limits, i.e., within the middle 70-80% of the full data set time range, performance is generally good (near fit scores were typically > 70%) and any variations in performance tended to be seen as smooth horizontal bands, rather than sloping and/or broadening bands indicating a relationship with data at that time in the supplied time-series rather than the influence of the investigated data property. Again, as expected, predictive power also typically decreased towards detection limits for magnitude predictions (indicated by red near fit regions in Figure S12). Here, positive and negative change detection limits both varied with input data gradient, but the range (positive limit – negative limit) was relatively consistent. This asymmetric behaviour, noted earlier for the base case which has a small negative gradient, was, however only apparent near the detection limit, and, again, away from the detection limit performance was generally good (near fit scores were > 70%) and most variations appeared to associate more strongly with input data rather than varying gradient. Detection limit did, however, improve significantly with reducing input data variance (see e.g. reduction in the size of undetected or empty-data region about 0% change, with reducing variance in Figure S13 bottom right. This suggests that pre-processing steps that further reduce variance could significantly improve detection limits. That said, decreased performance in the detection of larger magnitude changes (indicated by expanding yellow and red regions in Figures S13 and S14) strongly suggest that there are trade-offs involved and, e.g., over-smoothing of variance could limit the predictive power of subsequent break-point investigations. So, while this is encouraging in terms of the potential for further improvements in sensitivity, it is also an option that needs careful handling if robust quantification is the objective.

To investigate the effects of the local contribution isolation step, a second set of simulations were undertaken using the Kirkstall NO_2 time-series as the base case. While this dataset is not as directly representative of Headingley as the base case used for the previous simulations, this change was made because in this instance the

intention was to modify the properties of the EQ.4 formula used as part of the signal isolation step. As the methods used to generate the 'event-free' Headingley base case used in the previous simulations required the assumption of a form for EQ.4 when identifying events to remove, it was decided that modifications to EQ.4 were better investigated using a dataset that did not require prior event-removal. The results of the Kirkstall Road base case simulations are provided in the Supporting Information Figures S15 and S16. By comparison to Figures S6 and S7 trends, these are highly similar but arguably noisier, most likely reflecting the lower NO_2 levels at Kirkstall Road by comparison to Headingley. In addition, simulated events to the start and end of the time-series are more often not detected and data-related features (e.g. horizontal banding) were more pronounced in simulations using Kirkstall Road data as the base case.

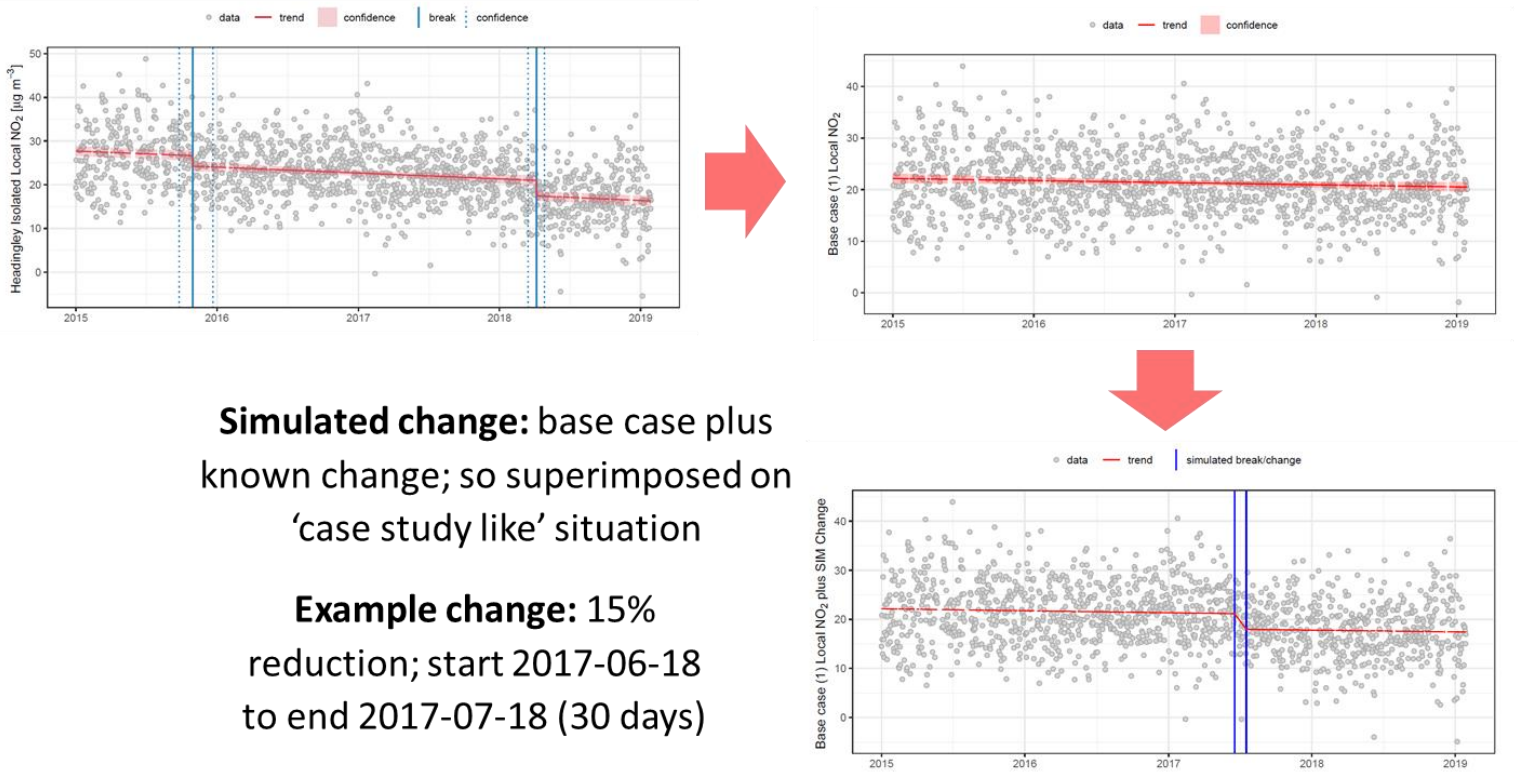
Simulation tests using meteorological data from different sources (Figure S17), clearly highlighted the benefits of using either the Ricardo WRF or the NOAA Integrated Surface Database/worldmet meteorological data rather than data from the nearest meteorological station. Arguably models that used the WRF data performed slightly better than those using worldmet data, but without more testing of more sites, this is probably best regarded as site/dataset-specific observation at this stage.

The influence of under/over-fitting was investigated by changing the number of knots used in the splines applied to the model inputs. As an example, Figure S18 shows the effect of applying knots in the range 3 to 30 to the combined wind speed and direction spline term used in EQ. 4. The number of knots in a GAM spline controls the 'wiggliness' of the fitting term, and model fit (agreement between modelled data and prediction of same data) generally increases with the number of knots. However, much like trained model performance on test-data (i.e., data not used to develop the model), break-point performance generally increases up to the point where the model becomes over-fitted and then performance starts to deteriorate. Here, it is important to acknowledge that the GAM models used in this study are by default subject to penalisation weighted regression[56,57], and that fits that actually applied the highest number of knots were only obtained if the penalisation term was removed. So, while this is not a strictly correct use of this particular model as its developer intended, it does demonstrate (1) the effect of non-linear fitting terms like knots in a spline, nodes in a neural network or trees in a forest, and (2) that fit statistics like the same-data regression coefficient of the isolation model do not necessarily provide a reliable measure of subsequent performance in break-point/segment tests. It is also worth noting that in this instance the best break-point/segment detection was observed for this spline term at about 15-19 knots, which is roughly where the penalisation weighted GAMs self-optimised.

The influence of the isolation model expression was also investigated, by addition and subtraction of input terms in EQ.4. Figure A19 uses a series of models, starting with a simple two-term (hour-of-day and day-of-year) model and building by addition to a seven-term (hour-of-day, day-of-year, background, wind speed/direction, air

temperature, day-of-week and month-of-year) model, to illustrate the effect of increasing isolation expression complexity on break-point/segment test outcomes. Here, as with knots and as expected, increasing the number of model inputs typically increased the isolation model fit (input data/model prediction agreement), but simulation once again demonstrated that isolation model fit was an unreliable indicator of sequent break-point/segment test performance. Unsurprisingly, the two-input (hour-of-day and day-of-year) performed poorest, but the addition of either a background term or a wind speed/direction term significantly increased break-point/segment detection rates and quantification accuracy. Optimal performance across the range of tested cases (1 to 100 day, +50% to -50% magnitude simulated change events) was seen for four-input (hour-of-day, day-of-year, background, wind speed/direction) models, although five-input (hour-of-day, day-of-year, background, wind speed/direction, air temperature) were only marginal poorer overall and arguably better at detection of smaller changes. Although beyond this point overall performance deteriorate further with each additional input, detection rates for smaller changes continued to improve, suggesting, albeit tentatively given the scale of current studies, that there are trade-offs here: more complex models may help to uncover smaller changes but they may also distort larger changes.

Base case: Headingley minus identified breaks; break-free dataset with very similar properties (variance, etc) to case study.



Simulated change: base case plus known change; so superimposed on 'case study like' situation

Example change: 15% reduction; start 2017-06-18 to end 2017-07-18 (30 days)

Figure S6: Schematic for data simulation using break-point subtracted Headingley time-series as representative base case for simulation testing.

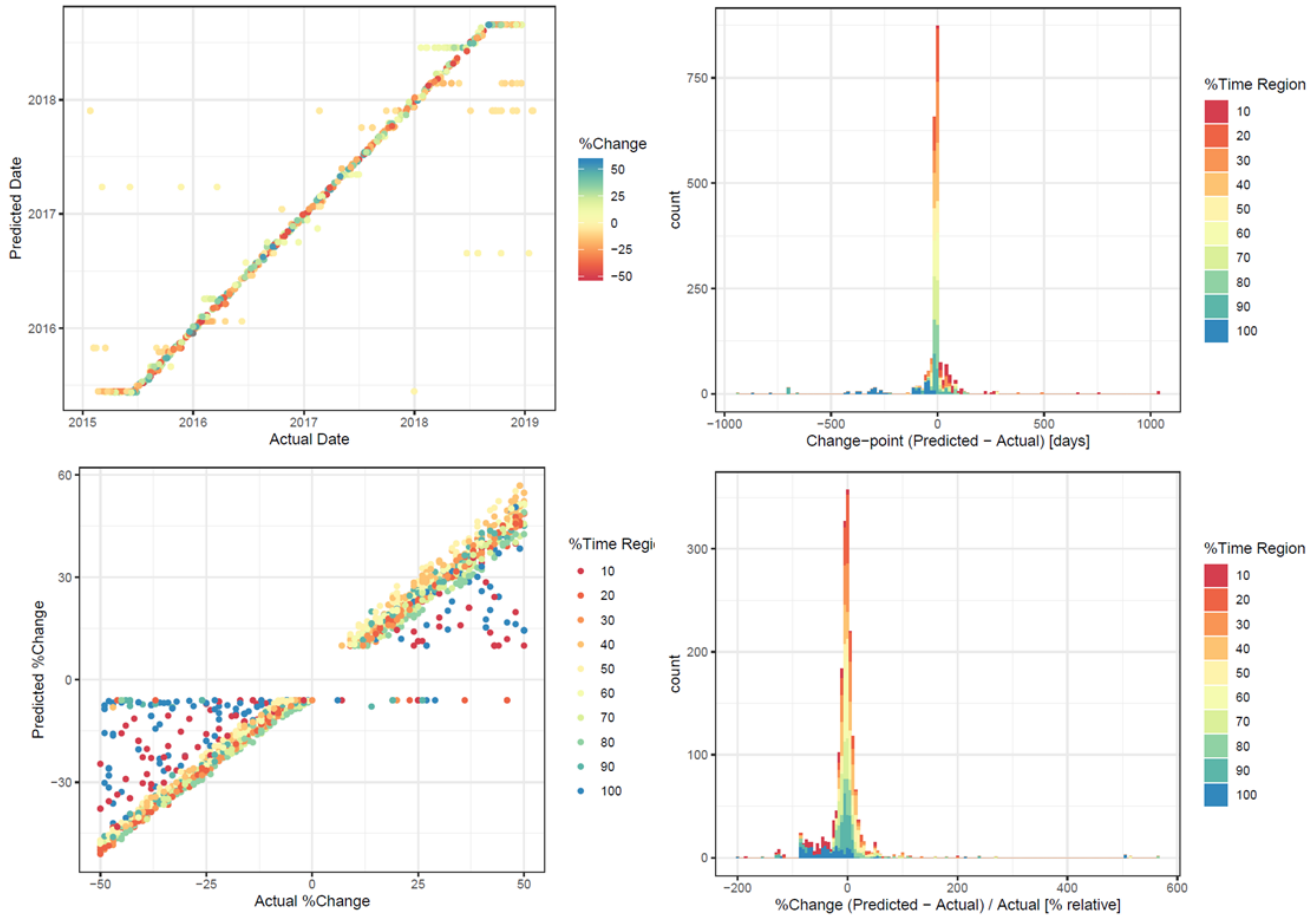
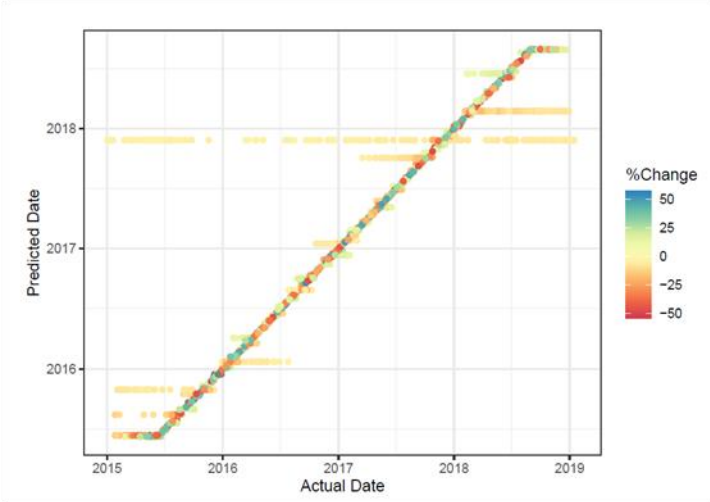


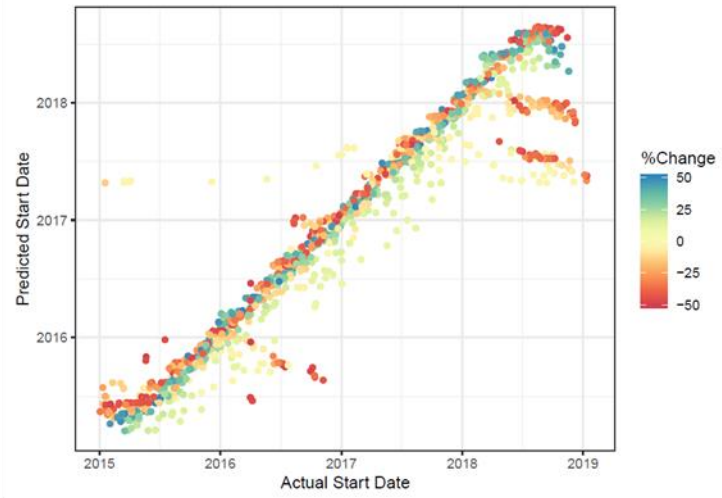
Figure S7: Simulation (n = 2,000) break-point testing instantaneous changes using ‘break-point-free’ Headingley data as base case.

Break-point detection: very accurate locating instant change…



[exceptions: at start and end (averaging window) and small changes (below detection limit)]

Change-segment detection: less certain of starts and ends of change-segments… but obviously these are much more sensitive to noise



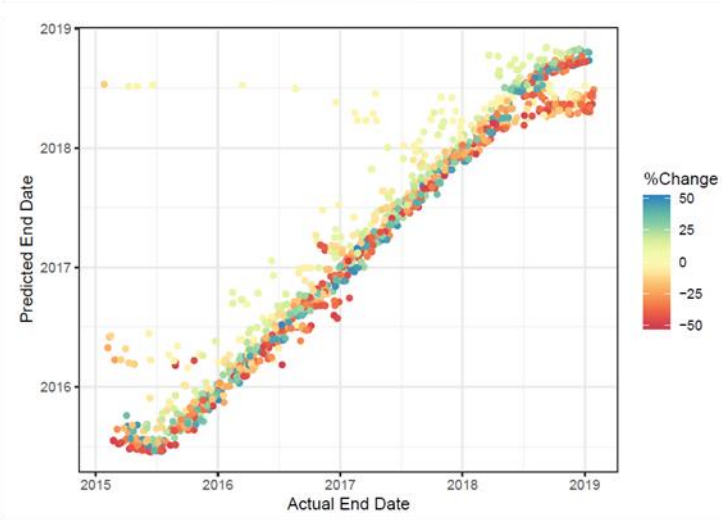


Figure S8: Comparison of break-point testing performance for instantaneous changes (left) and gradual changes (right), simulation n = 2,000 in both cases.

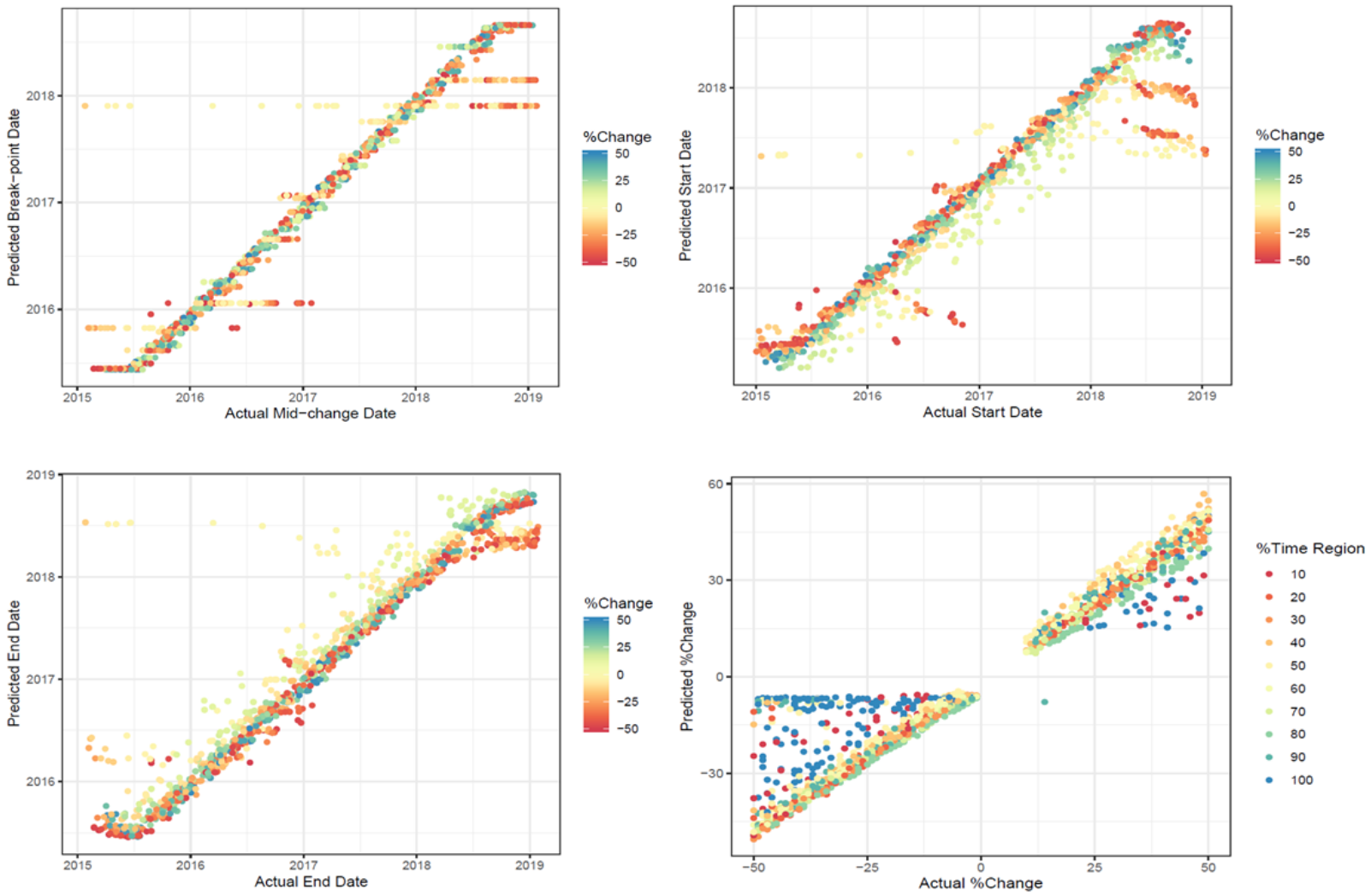


Figure S9: Simulation (n = 2,000) break-point testing gradual changes using 'break-point-free' Headingley data as base case.

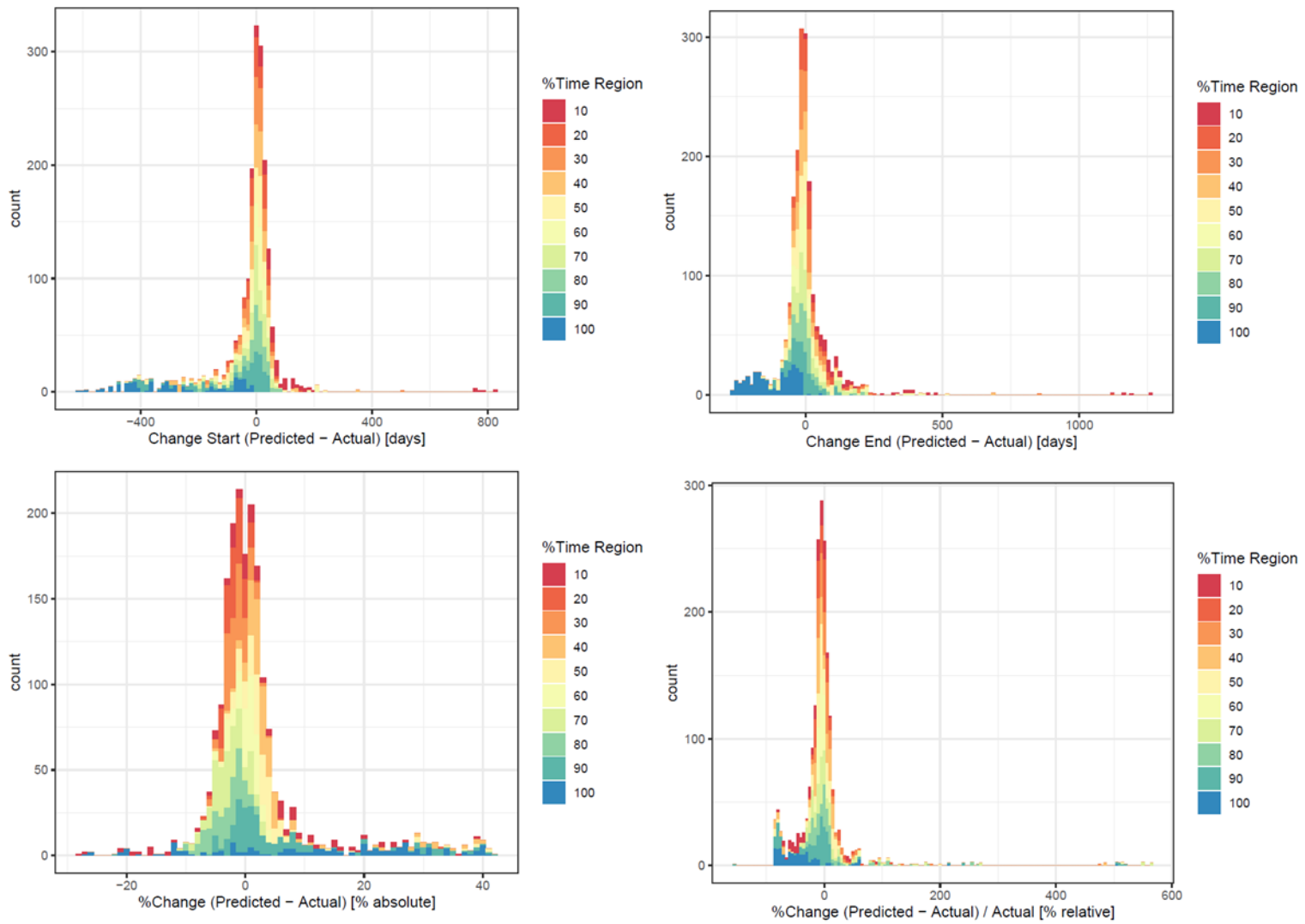


Figure S10: Simulation (n = 2,000) break-point testing gradual changes using ‘break-point-free’ Headingley data as base case.

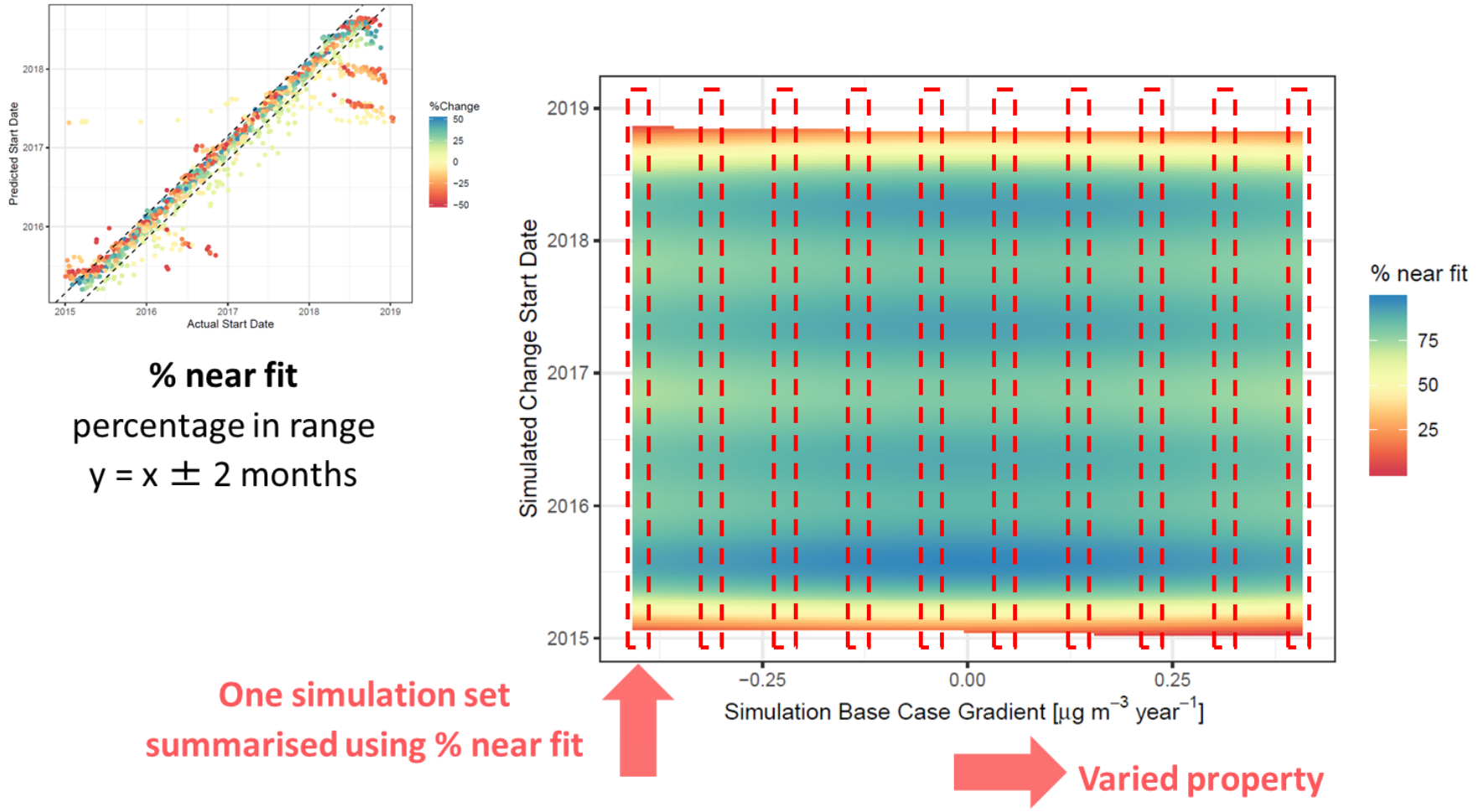


Figure S11: Simulation test schematic for the investigation base case data properties: For each simulation set (n = 2,000) the % near fit parameter was calculated as function of actual measurement, e.g. above actual start time in above example, left. Performance trend surfaces where then generated by repeating the process multiple times (typically about 10), each time modifying the investigated base case data property and fitting a surface to near-fit (y-axis) and property (x-axis), e.g. gradient in above example.

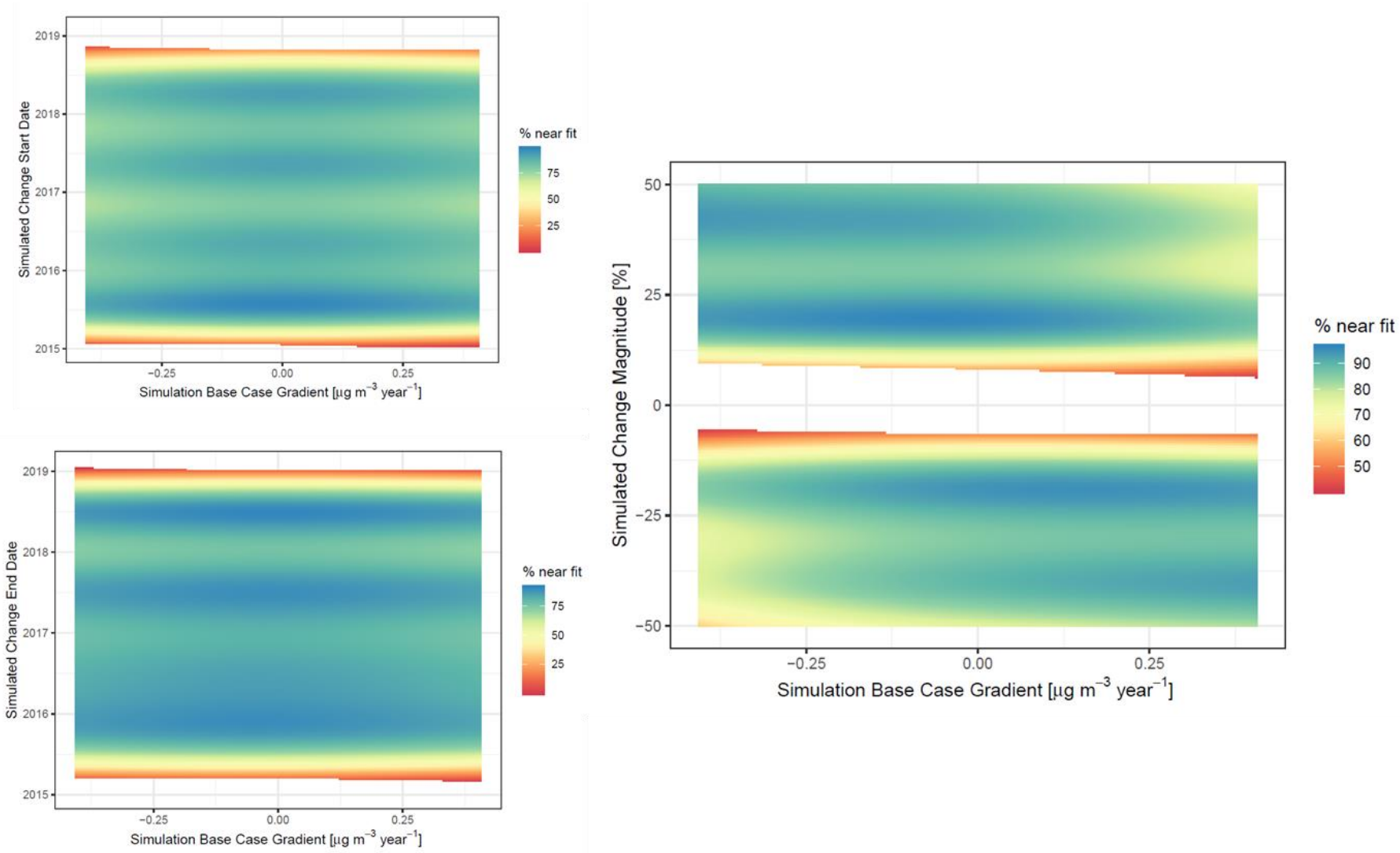


Figure S12: Simulation testing of effect of base case data properties: time-series gradient.

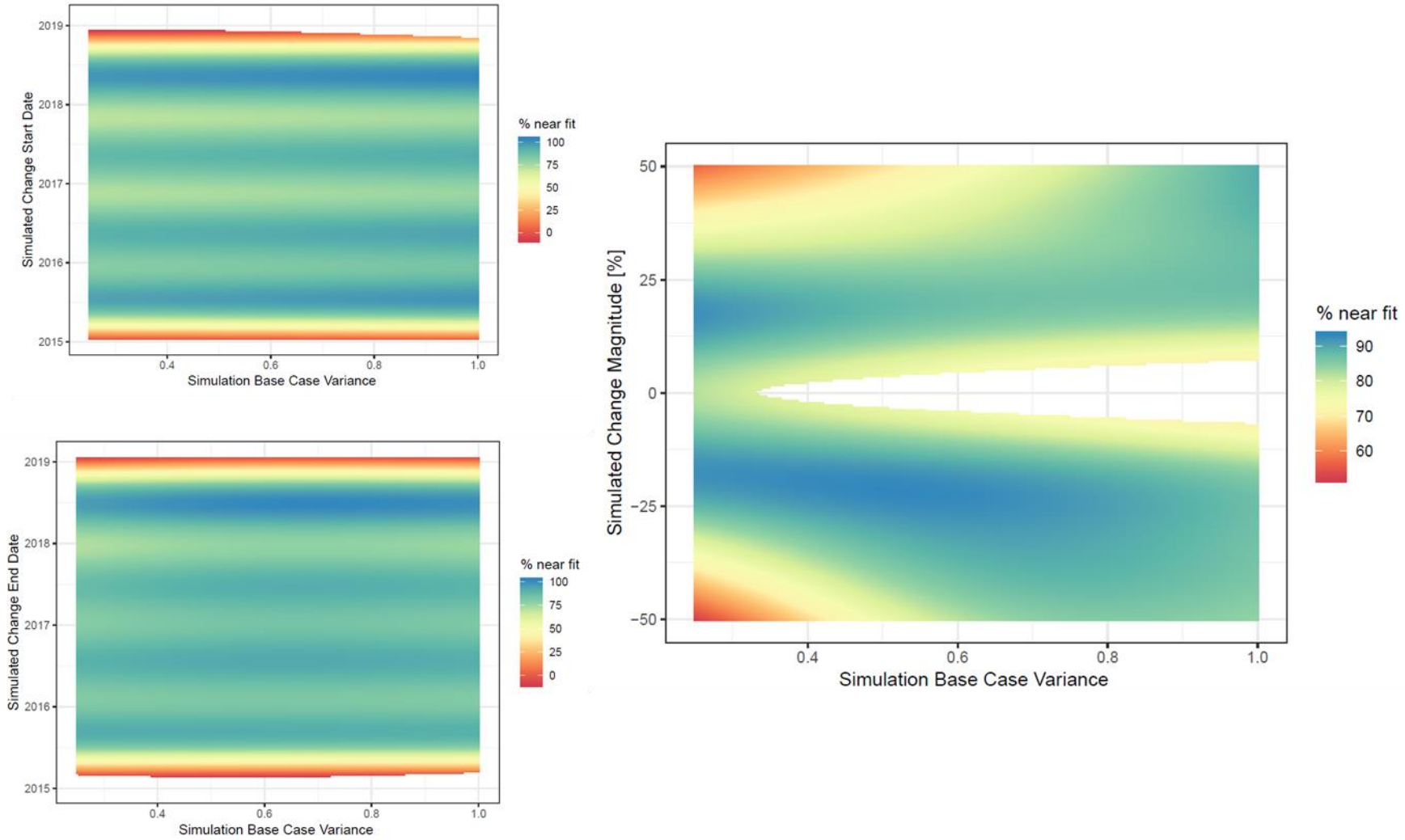


Figure S13: Simulation testing of effect of base case data properties: time-series variance.

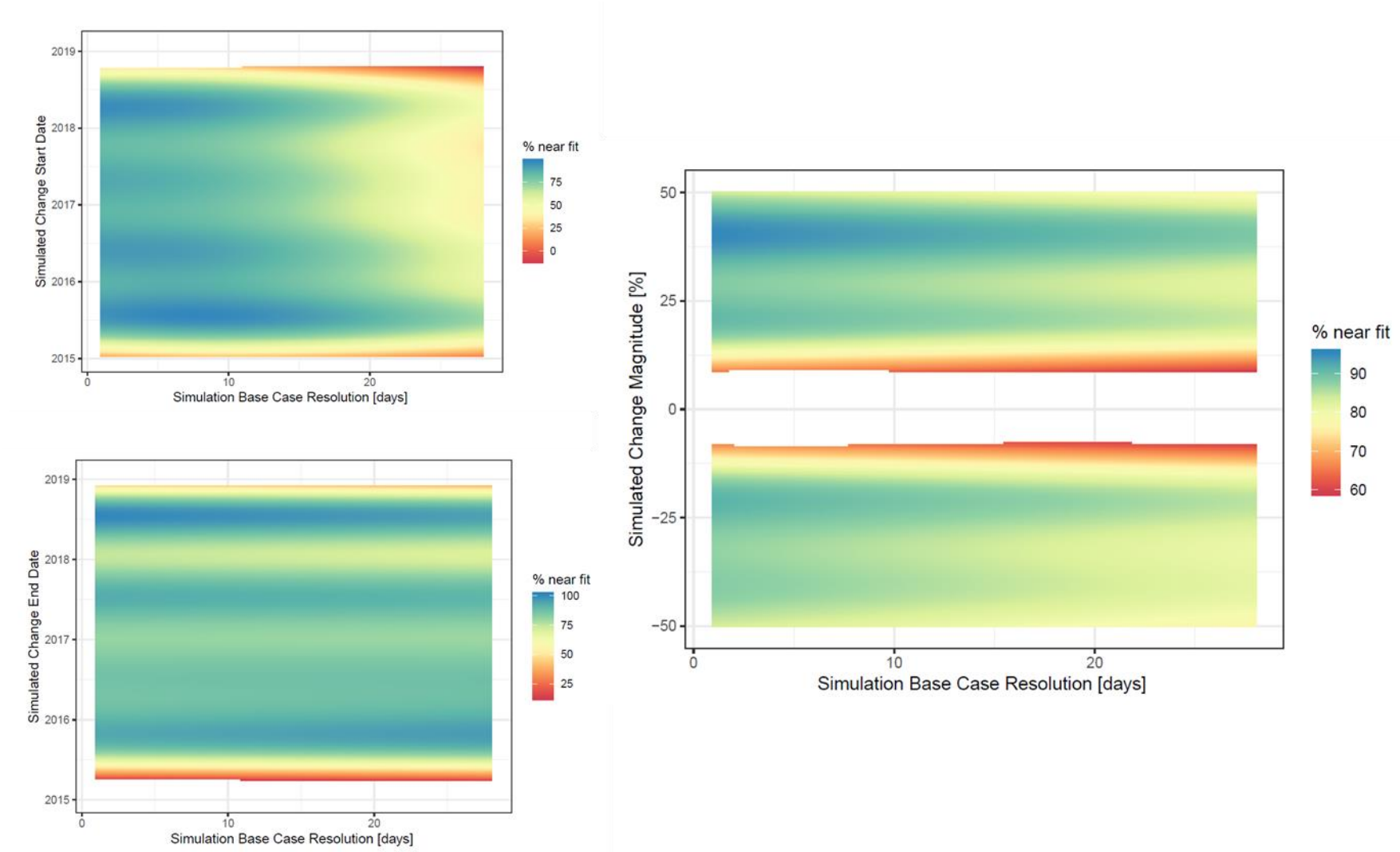


Figure S14: Simulation testing of effect of base case data properties: time-series time resolution.

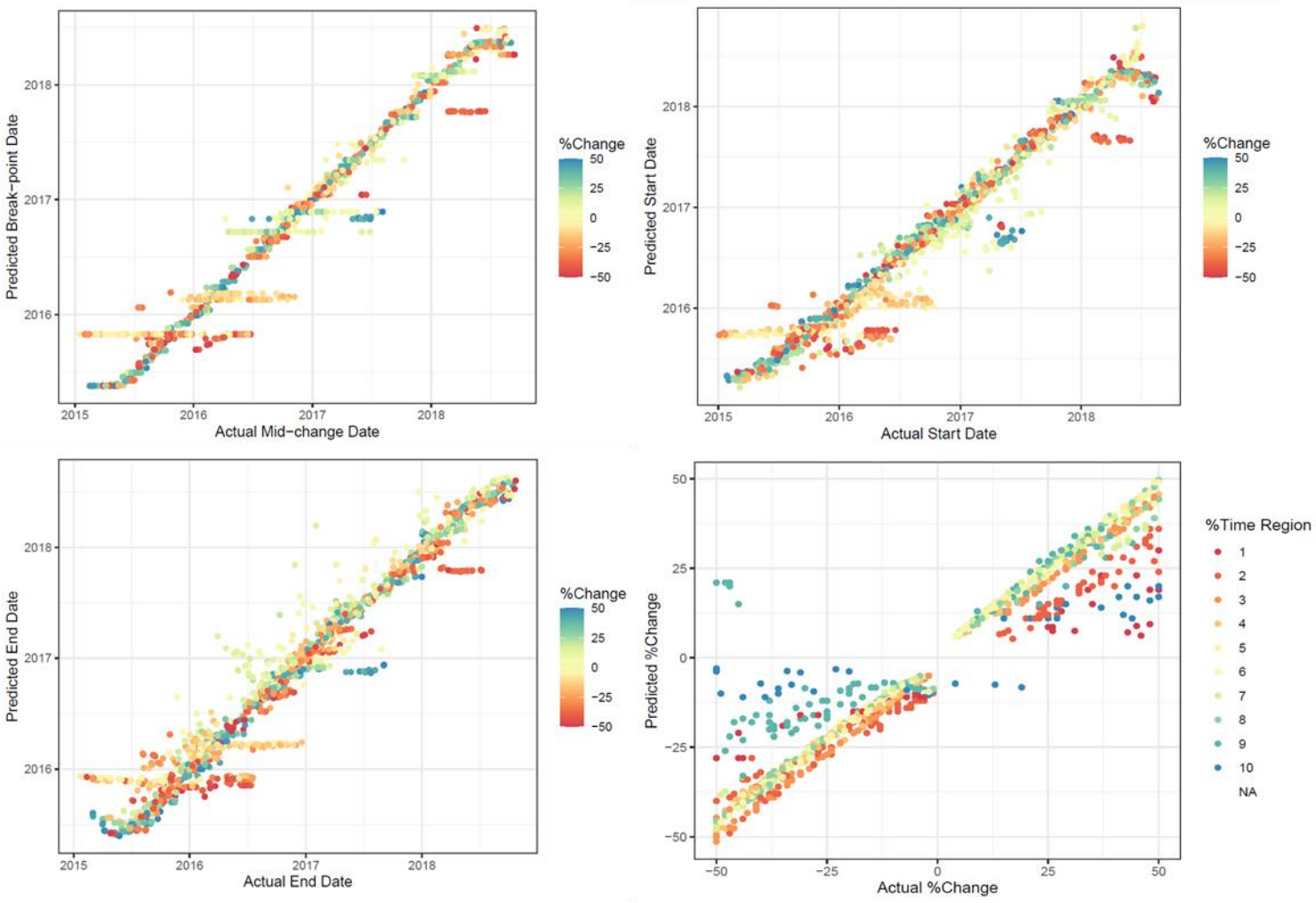


Figure S15: Simulation (n = 2,000) break-point testing gradual changes using Kirkstall Road data as base case.

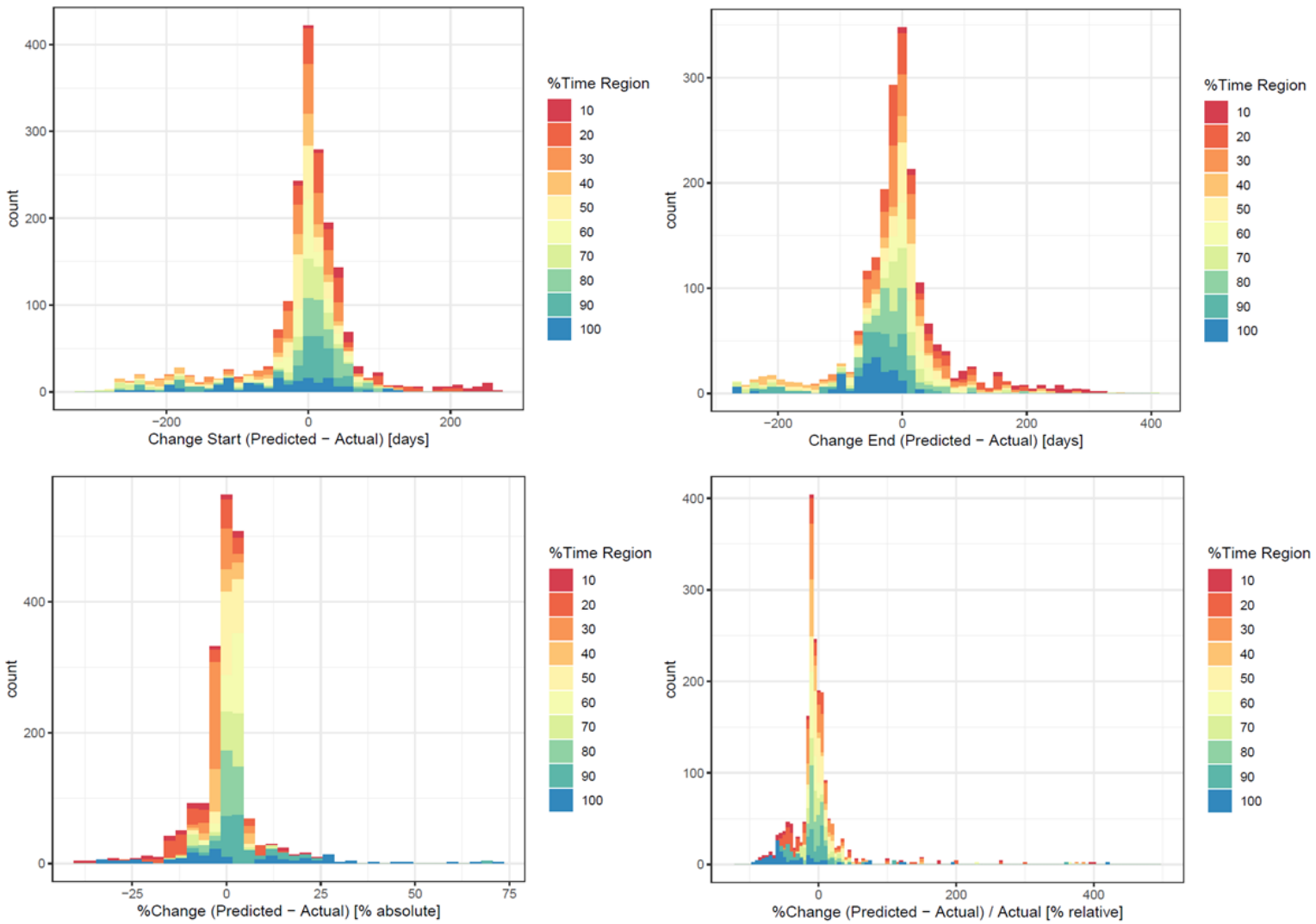


Figure S16: Simulation (n = 2,000) break-point testing gradual changes using Kirkstall Road data as base case.

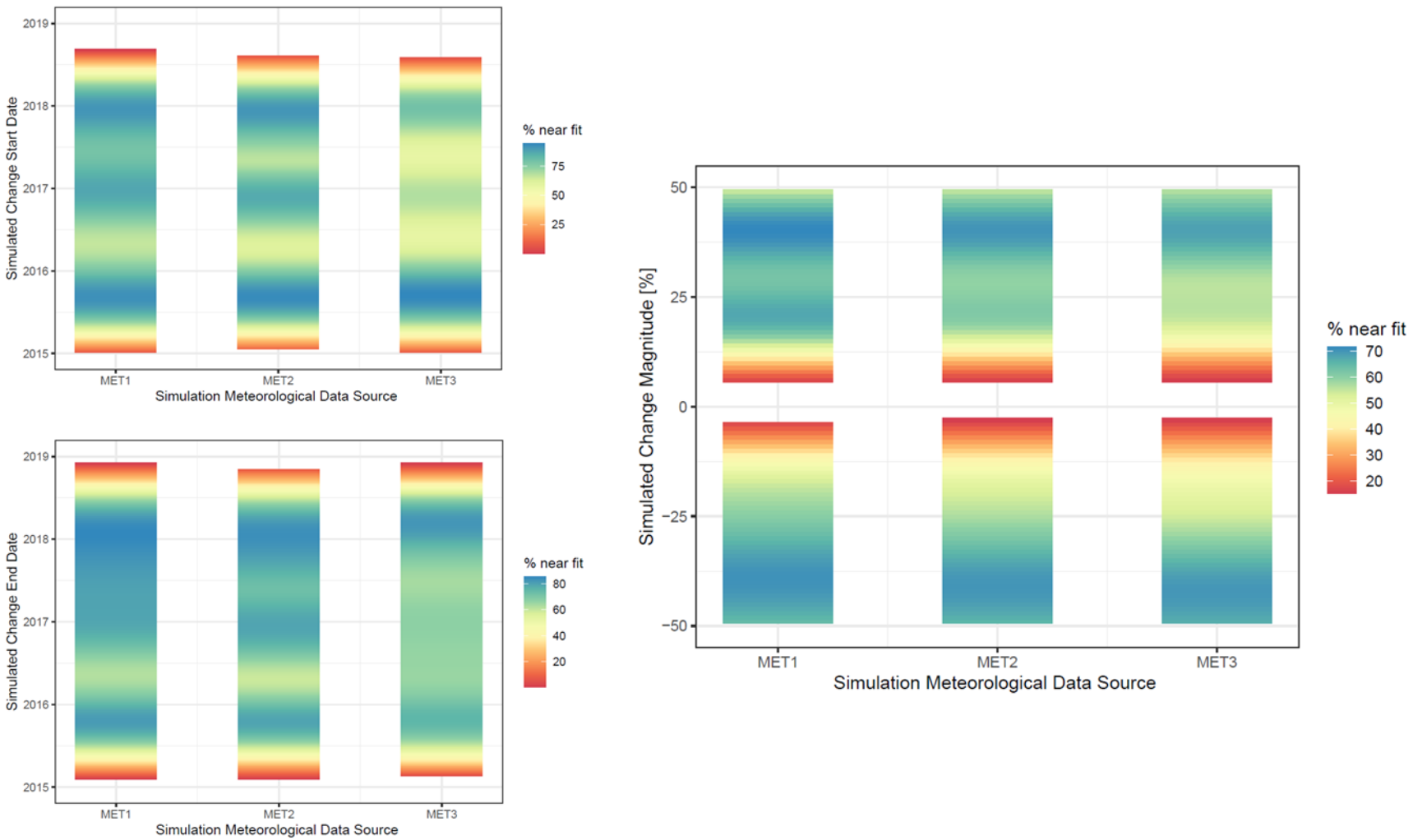


Figure S17: Simulation testing of effect of isolation model: meteorological data source. MET1 openair (WRF Ricardo) model (used in main study); MET2 Leeds Bradford MET via worldmet/NOAA; and, MET3 Leeds MET from local station. See also Table 1 and related discussion in main text.

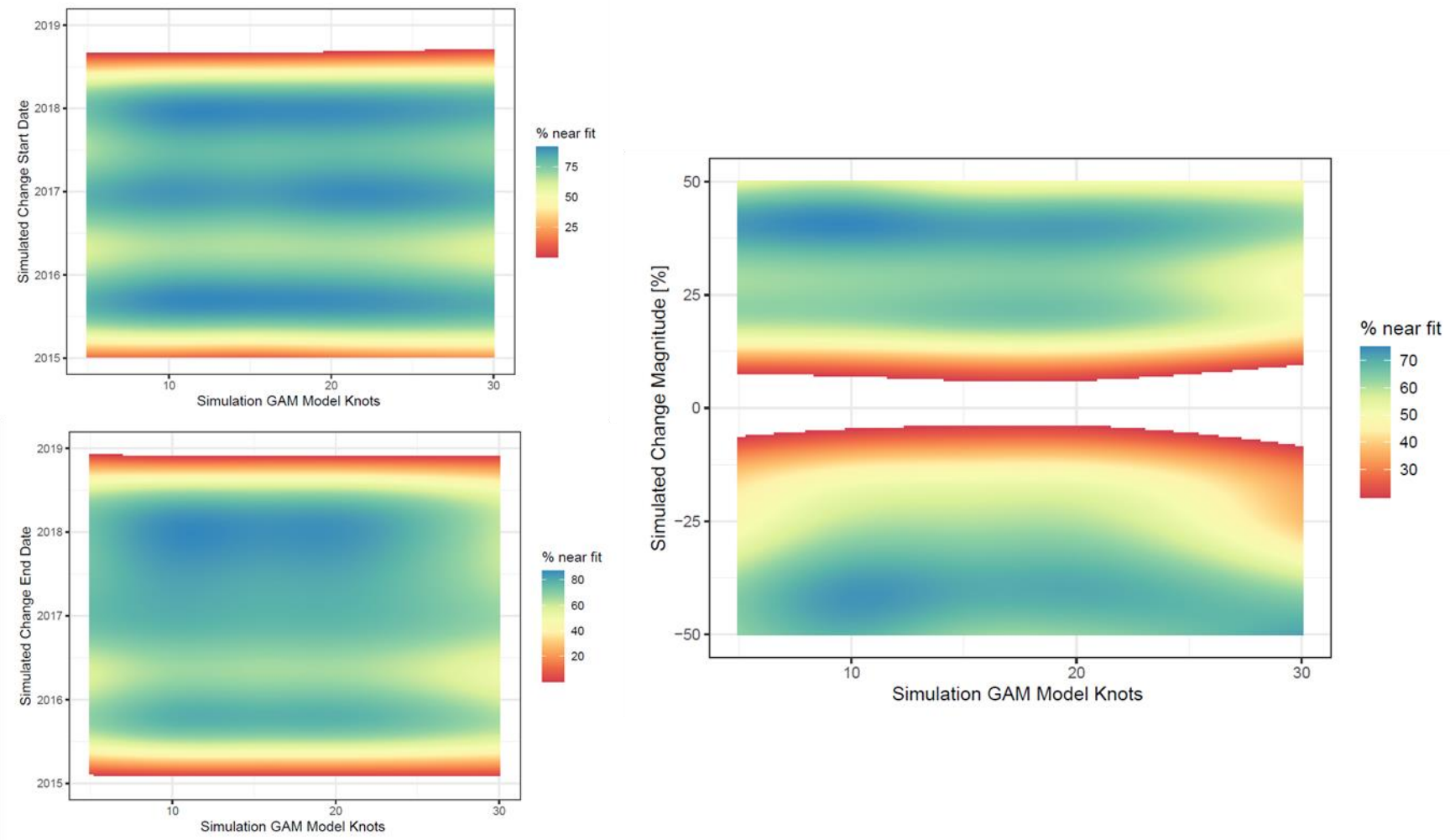


Figure S18: Simulation testing of effect of isolation model: changing number of knots applied by GAM model. The number of knots was forced when GAM fitting these simulations to demonstration effect of over/under-fitting. Without forcing the models typically self-optimised at 15-19 knots.

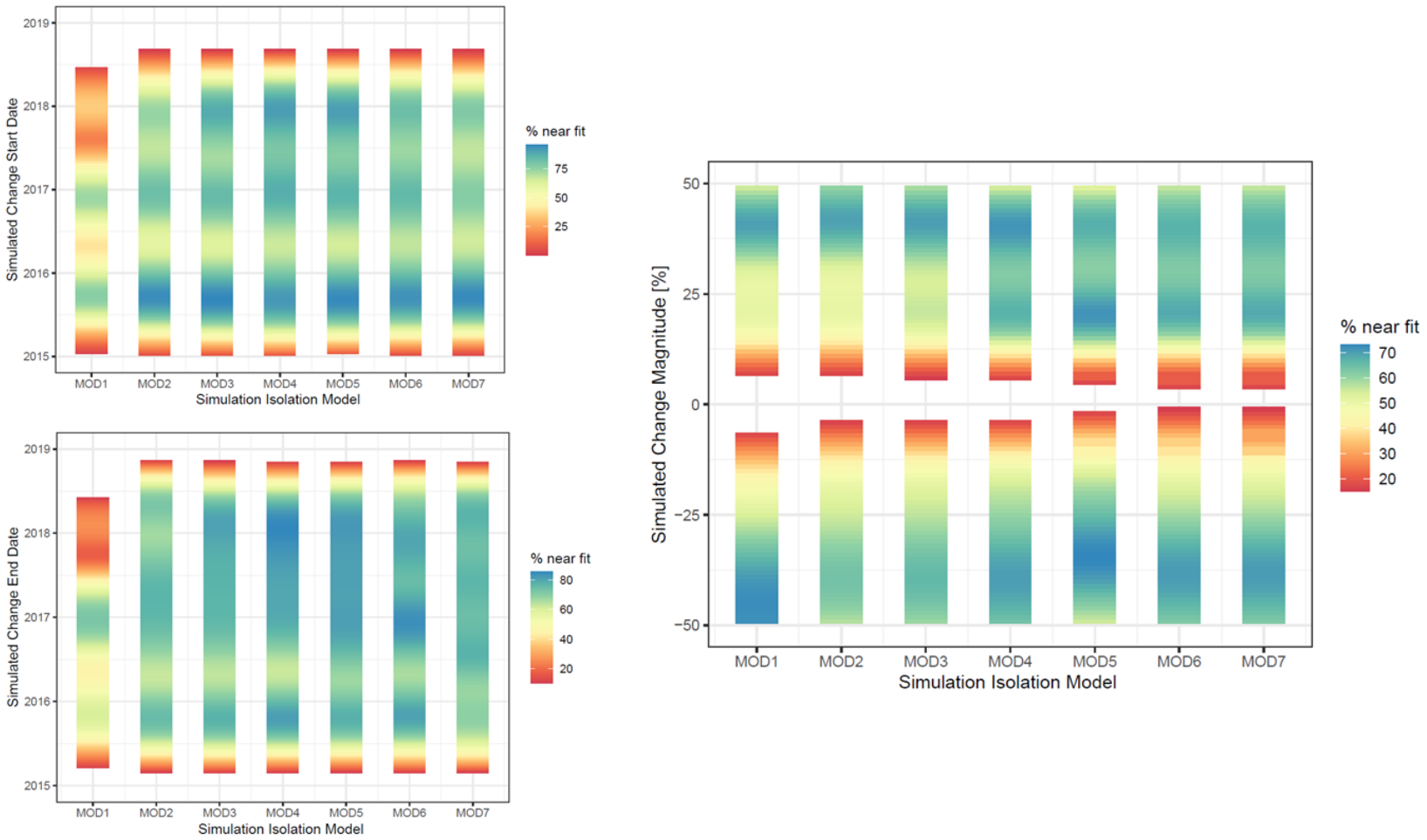


Figure S19: Simulation testing of effect of isolation model: changing model inputs. MOD 1 $[NO_2]_{site}$ = s(day.hour) + s(year.day); MOD 2 $[NO_2]_{site}$ = s(wind spd, dir) + s(day.hour) + s(year.day); MOD 3 $[NO_2]_{site}$ = s($[NO_2]_{BG}$) + s(day.hour) + s(year.day); MOD 4 $[NO_2]_{site}$ = s($[NO_2]_{BG}$) + s(wind spd, dir) + s(day.hour) + s(year.day) (model used in main study); MOD 5 $[NO_2]_{site}$ = s($[NO_2]_{BG}$) + s(wind spd, dir) + s(air temp) + s(day.hour) + s(year.day); MOD 6 $[NO_2]_{site}$ = s($[NO_2]_{BG}$) + s(wind spd, dir) + s(air temp) + s(day.hour) + s(year.day) + s(week.day); MOD 7 $[NO_2]_{site}$ = s($[NO_2]_{BG}$) + s(wind spd, dir) + s(air temp) + s(day.hour) + s(year.day) + s(week.day) + s(month.year).