Microbial diversity mediates the impact of air pollution on pneumococcal disease risk

Sophie Belman[1,2]; Cebile Lekhuleni[3,4]; Jackie Kleynhans[3,5]; Giovenale Moirano[1,9]; Daniela Lührsen[1]; Cristina Carnerero[1]; Susan Meiring[10]; Stephanie W. Lo[2]; Mignon du Plessis[3,4]; Anne von Gottberg[3,4,6,+] & Rachel Lowe[1,7,8,+]

1) Barcelona Supercomputing Center (BSC), Barcelona, Spain
2) Parasites and Microbes, Wellcome Sanger Institute, Hinxton, UK
3) Centre for Respiratory Diseases and Meningitis, National Institute for Communicable Diseases of the National Health Laboratory Service; Johannesburg, South Africa
4) Department of Clinical Microbiology and Infectious Diseases, School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
5) School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
6) Division of Medical Microbiology, Department of Pathology, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa
7) Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain
8) London School of Hygiene and Tropical Medicine, London, UK
9) Department of Medical Sciences, University of Turin, Italy
10) Division of Public Health Surveillance and Response, National Institute for Communicable Diseases of the National Health Laboratory Service; Johannesburg, South Africa

+ These authors contributed equally to this work

Abstract

Streptococcus pneumoniae (pneumococcus) remains a major global cause of invasive human disease with clear seasonal variation suggesting a role for environmental factors. The pneumococcus includes >100 serotypes and many lineages with different disease patterns. Interactions between microbial diversity, environmental factors, and invasive pneumococcal disease (IPD) remain unclear. We examined the effects of air pollution and meteorological factors on 59,017 IPD cases over 15 years, including 4,350 genome-sequenced isolates, using Bayesian spatio-temporal hierarchical models that link environmental exposures to bacterial diversity. Humidity and air pollution conferred the highest disease risk. In South Africa low humidity increased risk while high humidity was protective. Weekly $PM_{2.5}$ 50 μg/m³ was associated with 4% cumulative risk increase with diversity-specific lag structures suggesting partially vaccine-mitigated risk. $PM_{2.5}$ risk was greatest in adults ≥65 and varied by disease type. Understanding these relationships may refine vaccine-impact models and inform air pollution policy targets.

Introduction

Streptococcus pneumoniae (the pneumococcus) is asymptomatically carried in the human nasopharynx and can cause both non-invasive diseases (such as otitis media and sinusitis), and invasive pneumococcal disease (IPD) such as bacteraemia and meningitis, when the bacterium breaches the epithelium or enters the bloodstream[1,2]. The progression from colonization, to asymptomatic carriage, to invasive disease is mediated by a combination of environment, microbial virulence, and host immunity. The pneumococcus has hundreds of known genomic lineages with hundreds of years of diversity among them. Here we use Global Pneumococcal Sequence Clusters (GPSCs) to classify pneumococcal lineages. The pneumococcus is typically enclosed in a polysaccharide capsule of which there are more than 100 known serotypes. The seven-valent and 13-valent pneumococcal conjugate vaccines (PCV7 and PCV13) target seven and 13 serotypes respectively. This has resulted in serotype replacement by non-vaccine serotypes (NVTs) in the years following vaccine implementation globally[3–6].

Despite high vaccination coverage, IPD shows consistent seasonal spikes each year in part due to the extensive diversity of the pneumococcus. Efforts to reduce residual IPD burden is focused on adapting vaccination strategies to settings with varied sociodemographic contexts and pneumococcal colonization densities[7–10]. However, the role of air quality and meteorological factors, such as air pollution, temperature and humidity in pneumococcal colonization and disease dynamics is typically overlooked. Many geographic regions with the highest pneumococcal disease burden also have populations with extensive exposure to polluted air[14]. Understanding the sensitivity of pneumococcal dynamics to air quality and meteorological factors could allow vaccination strategy modeling and winter hospital capacity preparations to better adapt to different geographies and changing climates. Accounting for underlying pneumococcal diversity when studying the role of environmental factors in driving IPD transmission and incidence might be crucial given the known variability in virulence, carriage durations, and sociodemographic populations affected by different lineages and serotypes. By accounting for pneumococcal diversity we aim to additionally uncover the biological mechanisms behind environment-microbe interactions.

Here we focus on particulate matter <2.5μm in size ($PM_{2.5}$), and the associated risk for IPD while accounting for the underlying microbial diversity as well as additional environmental factors. We investigated the role of $PM_{2.5}$ in influencing IPD incidence using IPD case data spanning 2005 to 2023 (incorporating genomic data in models through 2019), from 531 hospitals in South Africa. This is a setting with pneumococcal colonization estimates ranging from around 40-60% (typically measured in under 5 year-olds)[15,16], and marked geographic and climatic diversity. Major public health interventions occurred during the study period (PCV7 introduction in 2009, PCV13 in 2011, and widespread

antiretroviral therapy (ART) for treatment of people living with HIV (PLHIV), since 2005). We include underlying pneumococcal phenotypic and genomic data to understand interactions with underlying microbial ecology, accounting for different serotypes and genomic lineages. We apply Bayesian hierarchical models to capture spatial and temporal variability, incorporate lagged relationships between environmental exposure and disease onset, and allow for stratification by age group and disease type. These models enable a better understanding of the interaction between pneumococcal populations and environmental factors which plays a role in IPD morbidity and mortality.

Results

A total of 59,017 identified IPD cases were included from 531 hospitals in South Africa between 2005-2023. The majority of cases were found in urban centers where the majority of the South African population resides (Figure 1A-B). These included bacteremia, meningitis, and other IPD types. The majority of cases occurred in people aged 15-64 years old (62%), followed by children under 6 years of age (23%). 4.8% of cases were from over-64 year olds. There was a well characterized decrease in serotypes that were targeted by PCVs. We observed NVT persistence (Figure 1C) and consistent seasonality in IPD cases. Cases peaked in colder months across the time period coinciding with the seasonality of $PM_{2.5}$ concentration and inversely correlating with relative humidity (Figure 1D)(Figure S1)(Table S1).

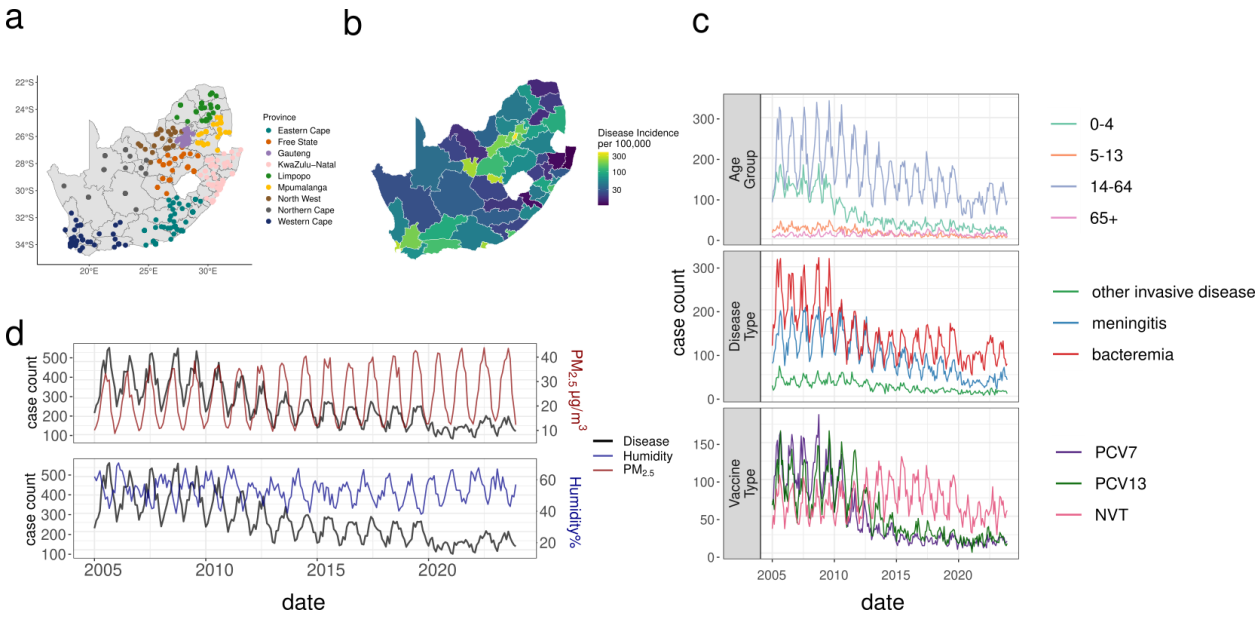

Figure 1. Summary of epidemiological and environmental datasets in South Africa. (a) Hospital locations (N=531) colored by province (N=9) (b) Log10 incidence rate per 100,000 from 2005-2023 by district (N=52) in South Africa (c) monthly case counts by age group, disease type (bacteremia, meningitis, other IPD types), vaccine type (PCV7 = seven-valent vaccine type serotypes; PCV13 = additional six serotypes in the 13-valent vaccine type; NVT = non-vaccine types), and serotype (colored by PCV7, PCV13, and NVT). (d) monthly pneumococcal disease case counts (black) with mean monthly particulate matter <2.5 µm concentration (red), and mean relative humidity (blue) across South Africa over 18 years; left axis is the monthly case counts, right axis is the concentration (µg/m^3) of particulate matter <2.5µm (upper panel) and percent relative humidity (lower panel).

Seasonal, interannual, and spatial effects

Our baseline model (M12) included spatial autocorrelation across districts, seasonal, and interannual random effects replicated across provinces; while also accounting for the pneumococcal vaccine implementation and population density (Table 4). We found similar effects for models spanning 2005-2023 and 2005-2019 but due to the dramatic decrease in IPD cases with the pandemic period we proceeded with models restricted to the 2005-2019 period throughout (Table S2-3). The baseline models were fitted to weekly invasive disease incidence across South Africa both at district (N=52) and province levels (N=9). The Gauteng and Western Cape Provinces had a higher risk of IPD than the national average. The relative risk (RR) derived from the spatial effect for Gauteng Province was 13.35 (95% credible intervals (cis) 4.19-44.34) and for Western Cape Province was 1.54 (95% cis 0.84–2.70), representing the area specific risk relative to the average across all provinces. The risk in Western Cape Province was largely driven by the increased risk in the Cape Town district alone (relative to all district average)(RR: 10.16; 95% cis 7.89-13.09). All districts in Gauteng Province exhibited elevated disease risk relative to the national average (RR>1). Some of the exacerbated risk in the main cities may reflect travel to hospitals from more rural areas (Figure 2A)(Figure S1C). Across the other provinces, districts which host major cities and transport hubs had twice the risk of IPD compared to average (Figure 2A)(Table S3). There was a seasonal effect with an increased disease risk between the months of May and October with peak risk in August (RR: 1.37; 95% cis 1.33-1.41). In the interannual effect we found a reduction in risk over time even while accounting for vaccination (Figure 2C). This may be attributed to uptake of ART (Figure S1D).

There was a correlated increase in coverage for ART and vaccination across the study period (Pearson R^2 = 0.84) and their inclusion in the models reduced interannual effect variance (Figure S1D)(Figure S2A). We found no effect from ART or PCV on NVT serotypes alone. Interestingly, models estimating NVT disease (in this high HIV burden setting)[17] had substantial impacts on IPD when we include years 2020-2023 (Figure S2C-D). This may reflect the increase in cases among adults ages 18-64 after the SARS-CoV-2 reductions in 2020 and waning vaccine effect since the rapid uptake (Figure S2E)[18–20]. Due to the established direct relationship between pneumococcal vaccination and IPD, and the nuanced impact of ART, we included the PCV periods as the intervention-covariate in the baseline model (Figure 2D)(Table S4).

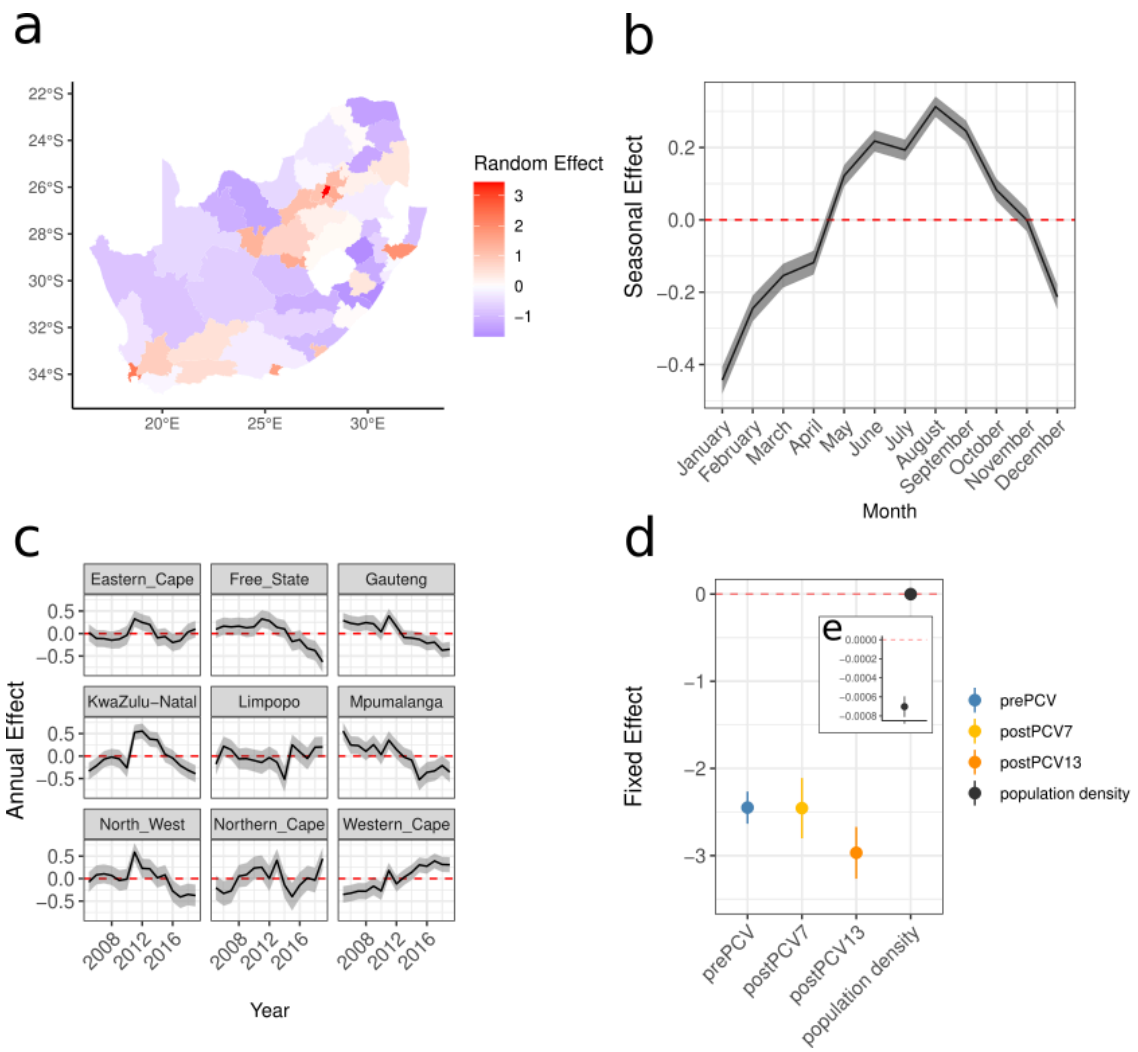

Figure 2. Evaluating spatial, seasonal, and interannual effects for a base model and comparing interventions (a-c) Random effects for the base model including spatial, seasonal, and interannual effects as well as accounting for implementation of the vaccine in 2009, 2011, and population density (a) spatially across 52 districts where a positive effect (more cases than expected on average) is red and a negative effect is blue, (b) seasonally across the 12 months of the year, the red line denotes effect size of 0, (c) across each of 2005-2019 included in the model replicated across the 9 provinces of South Africa (d) fixed effects for each vaccine period prePCV (2005-2008) (blue), postPCV7 (2009-2011) (yellow), and postPCV13 (2012-2019) (orange), and for the population density (black) (e) an inset of the population density effect.

Contribution of environmental factors to pneumococcal disease risk

We fitted the same baseline model as above, iteratively, with univariable environmental factors to estimate the risk of IPD across an 8-week time period from environmental exposure (M12, Table 4). We used distributed lagged non-linear models (DLNMs) to incorporate environmental factors (Figure S3). Model selection based on WAIC from univariable models, indicated that relative humidity, precipitation, $PM_{2.5}$ and particulate matter <10µm (PM_{10}), sulfur dioxide (SO_2), and maximum temperature had the best fit. We estimated the risk of disease above the 2016-2029 NAAQS thresholds[21] for $PM_{2.5}$ (40 µg/m^3), PM_{10} (75 µg/m^3), and SO_2 (125ug/m³).

Humidity and precipitation were negatively associated with IPD risk, whereas $PM_{2.5}$, PM_{10}, and SO_2 were positively associated with IPD risk (cumulative 8-week effect). Maximum temperatures below 20°C and above 30°C were negatively associated with IPD, and maximum temperatures from 25-30°C were associated with an acute increased risk (in the first 1-2 weeks). These patterns persisted when models were fitted to Gauteng alone (Figure S4)(Figure S5)(Table S5).

To understand the effect of $PM_{2.5}$ and PM_{10}, together with other environmental factors, we also explored the multiplicative effects of including multiple environmental variables in a single model. There was a higher risk of IPD between 35% and 47% relative humidity, which remained consistent after accounting for $PM_{2.5}$ or PM_{10} as a multiplicative effect. Relative humidity of 39.8% had the highest IPD risk (RR = 1.05; 95% CI: 1.03–1.07)(Figure 3C). There was a strong protective effect of high relative humidity with a 25% decreased risk of pneumococcal disease at 74.8% relative humidity (RR: 0.77; cis 0.73-0.80) (Figure 3E). The risks from relative humidity waned after about four weeks. The acute disease risk with low absolute humidity (which in contrast with relative humidity does not account for the temperature-dependent variation in the air's capacity to retain water) was strongly attenuated by $PM_{2.5}$ and $PM_{10,}$ however, with a 2-week lag time the risk from low absolute humidity was 1.23 (95% cis 1.16-1.31)(Figure S6)(Table S5).

With a weekly average of 50 µg/m³ of $PM_{2.5}$ there was a 4% 8-week cumulative increased risk of IPD (RR 1.04 [95% cis 1.03-1.5]), but accounting for relative humidity decreased that risk to 3.6%. Strikingly, in this model there was a 21% increased IPD risk (1.21: 95% cis 1.16-1.27) with a weekly $PM_{2.5}$ of 100 µg/m³ (Figure 3A)(Figure S6A)(Table S5). The effects of $PM_{2.5}$ were immediate but the maximum risk was in the fourth week after exposure. The City of Johannesburg regularly exceeded weekly averages of 100 µg/m³ of $PM_{2.5}$ (32 weeks in 2023) and across South Africa the weekly $PM_{2.5}$ concentration was at least 50 µg/m³ in about 23 districts a year (about 14 weeks per district)(Figure S7)(Table S5). In 2019 Gauteng Province had a mean $PM_{2.5}$ concentration of 117 µg/m³ and 718 observed IPD cases, and the population attributable fraction, based on weekly $PM_{2.5}$ concentrations (accounting for relative humidity), and given causal assumptions in our modelled exposure–response relationship, was 2.06% (14.8 IPD

cases) meaning if $PM_{2.5}$ had been maintained at or below 40 µg/m³ 14.8 cases could theoretically have been averted; if this were reduced further, to the more stringent WHO threshold of 15 µg/m³, 24.3 cases (3.39%) would be averted. A weekly PM_{10} concentration of 100 µg/m³ increased the cumulative 8-week risk of disease 6.5% (RR 1.07: 95% cis 1.05-1.08) and similarly to SO_2 the effect was lagged with an increased risk of disease only 3-8 weeks after exposure (Figure S6).

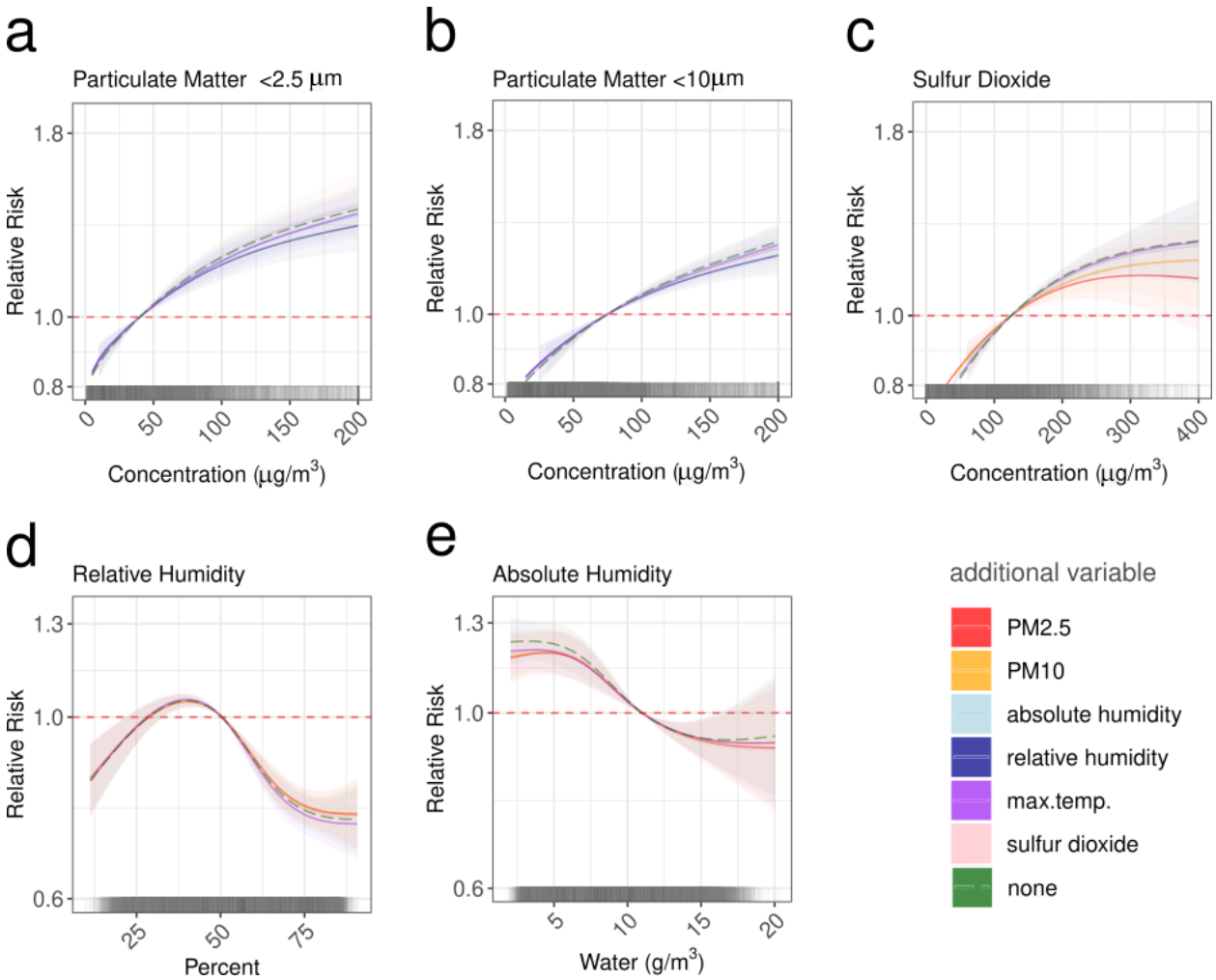

Figure 3. Cumulative relative risk across 8-weeks of lags for $PM_{2.5}$, PM_{10} SO_2, relative, and absolute humidity with additional multiplicative variables. In the DLNM the humidity variable cross basis has 3-degrees of freedom (df), while the pollution variables have 2. The additional multiplicative effects are included as fixed effects except the humidity variables which are non-linear with 5 cuts. Each plot includes the exposure-response curve for each lag-week colored by the multiplicative variable for (a) $PM_{2.5}$ (models for none, absh, hurs, tasmax, SO_2) (b) PM_{10} (models for none, absh, hurs, tasmax, so2), (c) SO_2 (models for none, $PM_{2.5}$, $PM_{10,}$ absh, hurs, tasmax) (d) % relative humidity (models for none, $PM_{2.5}$, PM_{10}, SO_2, tasmax) , and (e) absolute humidity (models for none, $PM_{2.5}$, PM_{10}, $SO_{2,}$ tasmax). This model is fit to weekly district level data from 2005-2019 across South Africa. At the bottom of each is a rug plot with the distribution of the main variables.

Risk from $PM_{2.5}$ varies by age, disease, and pneumococcal serotype

In the same environmental model formulations as above, we stratified outcome by particular demographic groups, disease outcomes, or pneumococcal serotypes. We ran each model with univariable DLNMs for $PM_{2.5}$, PM_{10}, SO_2, and relative humidity. Model performance varied but $PM_{2.5}$ and relative humidity had consistently good overall model performance (Table S5).

Age and disease outcome with $PM_{2.5}$

The highest risk from $PM_{2.5}$ was for those aged ≥65 years, with a cumulative RR across 8 weeks, and 50 μg/m^3 weekly average $PM_{2.5}$ exposure was 1.16 (cis 1.10-1.22)(Figure 4B). The relative risk ratio between each subset model and the main model (all IPD) translated to a relative risk ratio of 1.12 (95% cis 1.07-1.17) indicating a 12% higher risk for this age group at this $PM_{2.5}$ concentration (Figure 4A). Those between the ages of 15-64 had a more acute impact with a RR>1 in the week of high $PM_{2.5}$ concentrations (lag 0) while all other demographic groups showed similar lag patterns with the highest risk 3-4 weeks after exposure. This may reflect the higher prevalence of HIV among this age group[17]. The risk for those ≤5 years old was not negligible. For children ≤5 years, at 50 μg/m^3 weekly average there was a 8-week cumulative RR of 1.03 (cis 1.01-1.05), and for females of all ages it was 1.04 (cis 1.02-1.05). While similar to the average risk of IPD across the population it translates to 3% more children with IPD at high pollution levels and highlights that $PM_{2.5}$ does not only pose an increased disease risk for adults. (Figure S8). Models for particular disease outcomes (bacteremia, meningitis, or other invasive disease) had similar risk patterns. However, there was a higher cumulative risk (RR 1.12: 95% cis 1.08-1.16) and a more acute risk (>1 at early lag weeks) for other invasive disease types (Figure 4C)(Figure S8). The risk with a 50 μg/m^3 weekly average was higher for meningitis (RR 1.05: 95% cis 1.03-1.07) than for bacteremia (RR 1.02: 95% cis 1.00-1.04)(Figure 4C), possibly indicating different mechanisms of effect across IPD types.

Serotype and vaccine type with $PM_{2.5}$

The cumulative risk from $PM_{2.5}$ across all vaccine types was similar across the 8 weeks, there was a varied lag pattern with an immediate effect for NVTs and PCV7 serotypes but a longer time to effect for the additional serotypes in PCV13 (Figure 4D-E). These effects persisted when we stratified by pre- and post-PCV periods implying a difference in mechanism not just changing serotype proportions (Figure S9). When further stratified by specific serotypes variable underlying dynamics emerged. With an exposure of 50 μg/m^3 of $PM_{2.5}$ weekly we found a relative risk ratio of >1 as compared to on average in the population for PCV7 serotypes 14 and 23F, and it trended higher for NVT serotype 8 and PCV13 serotype 19A (Figure 4A)(Figure 4E)(Figure S10A-B). We found acute risks from $PM_{2.5}$ for NVT serotype 8 (94% GPSC3 in this dataset), and PCV13 serotypes 23F (68% GPSC14) and serotype 4 (58% GPSC70). These are all known invasive types. The 8-week cumulative risk for serotype 14 disease with 50 μg/m^3 average

weekly $PM_{2.5}$ exposure was greater than on average in the population (RR ratio: 1.08 95% cis 1.04-1.12)(Figure 4A). Serotype 14 is targeted by PCV13 and in South Africa predominantly is represented across several different lineages including the lineages GPSC10 (49%), GPSC18 (25%), and GPSC9 (20%)(Figure S10C). The predominantly lagged effect of $PM_{2.5}$ on vaccine serotypes demonstrates a protective effect from the vaccine against some of the immediate risks conferred from high levels of $PM_{2.5}$ for IPD.

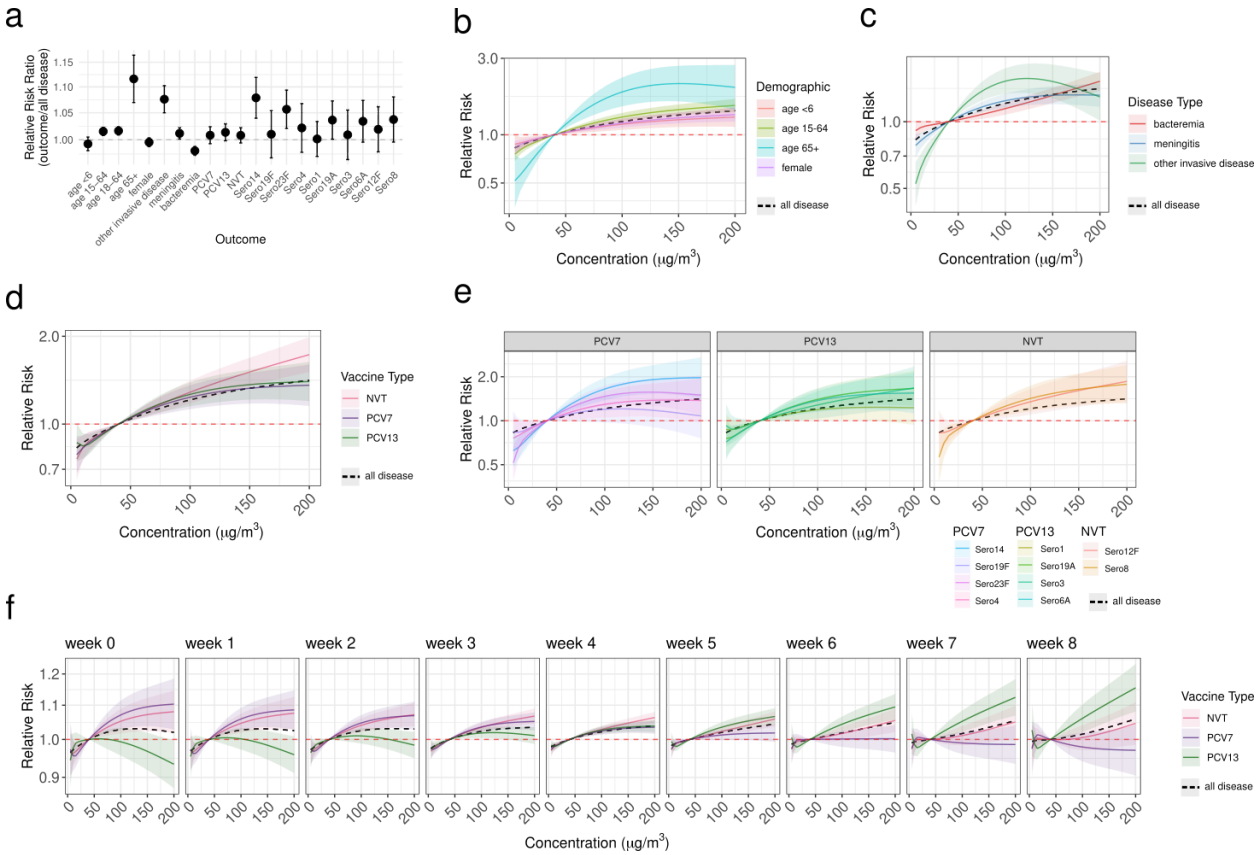

Figure 4. Risk of specific sub-outcomes of invasive pneumococcal disease given weekly averages of particulate matter <2.5μm ($PM_{2.5}$) in a baseline model including population density and vaccination period together with random effects at the district level from 2005-2019. (a) reporting the relative risk ratio of each specific model compared to the model including all disease at 50 μg/m^3 (exceeding the NAAQS threshold for $PM_{2.5}$ by 10 μg/m^3) (b) across age groups including age ≤5 (red), age 15-64 (green), and age 65 and older (blue), and including only females (purple), (c) the relative risk of $PM_{2.5}$ across disease types for bacteremia (red), meningitis (blue), and other invasive disease types (green), (d) models for disease caused by serotypes targeted by PCV7 (purple), additional 6 serotypes in PCV13 (green), and NVT (red) (e) and fitting models to disease caused by the particular serotypes in these groups faceted by vaccine type and colored by serotype. (f) the relative risk with a 0 to 8 week lag time from $PM_{2.5}$ to disease for each vaccine type group (pcv7=purple, pcv13=green, nvt=red). All models also include the RR from a model including all disease cases (black dashed).

Timing of air quality effects is mediated by lineage diversity

We estimated the impact of particular GPSC-lineage prevalence on the $PM_{2.5}$ effect on IPD at the province level, accounting for a 8-week lag time from the years 2005-2019. In these models we include IPD as the outcome but to incorporate genomic lineages into the model (N=4350) we included the proportion sequenced at each space-time increment as an effect modifier (Figure S11). We interacted the crossbasis for each environmental factor with the proportion of each dominant genomic lineage (N>=90 as well as GPSC8) among sequenced isolates (Figure S12)(Table S1). Models with GPSC41 and GPSC8 had worse goodness-of-fit metrics than the model which excluded lineage diversity interaction effects (Figure S13)(Table S5).

We explored the risk of IPD given the GPSC prevalence across concentrations of $PM_{2.5}$ (where prevalence is across quantiles of GPSC proportions over the entire time period; low=0.1 and high=0.9)(Figure 5A)(Table S6). While the 8-week cumulative effect did not vary with lineage diversity there was variation in the risk across lags and GPSCs (Figure S13A-B). There was an acute effect for GPSC21 whereby the risk of IPD was elevated in time-space increment with high prevalence of these lineages. The cumulative effect with a GPSC21 interaction also had different exposure-response curves between high and low prevalence (Figure 5C). We found an early effect whereby locations with a high prevalence of GPSC21 had an acute risk of IPD immediately following high concentrations of $PM_{2.5}$, whereas times and locations with low prevalence of GPSC21 have an increased risk at later lag times (Figure 5C). In this dataset GPSC21 comprises serotype 19F (n=90) with one instance of 19A in 2005. Serotype 19F is targeted by PCV7 and is characterized in South Africa by an initial GPSC1-lineage (n=114) which decreased post-PCV when GPSC21-19F became predominant (Figure 5B)[22]. When we re-run the GPSC21 interaction models splitting by pre- and post-PCV period we see the effect persisting in the post-PCV period with poor model performance due to low-N in the pre-PCV period (Figure S14). While further investigation is required to understand the mechanism by which GPSC21-19F is causing invasive disease in the short-lag period following high levels of $PM_{2.5}$, it is clear that there are some differences in environmental effects given the circulating lineage diversity.

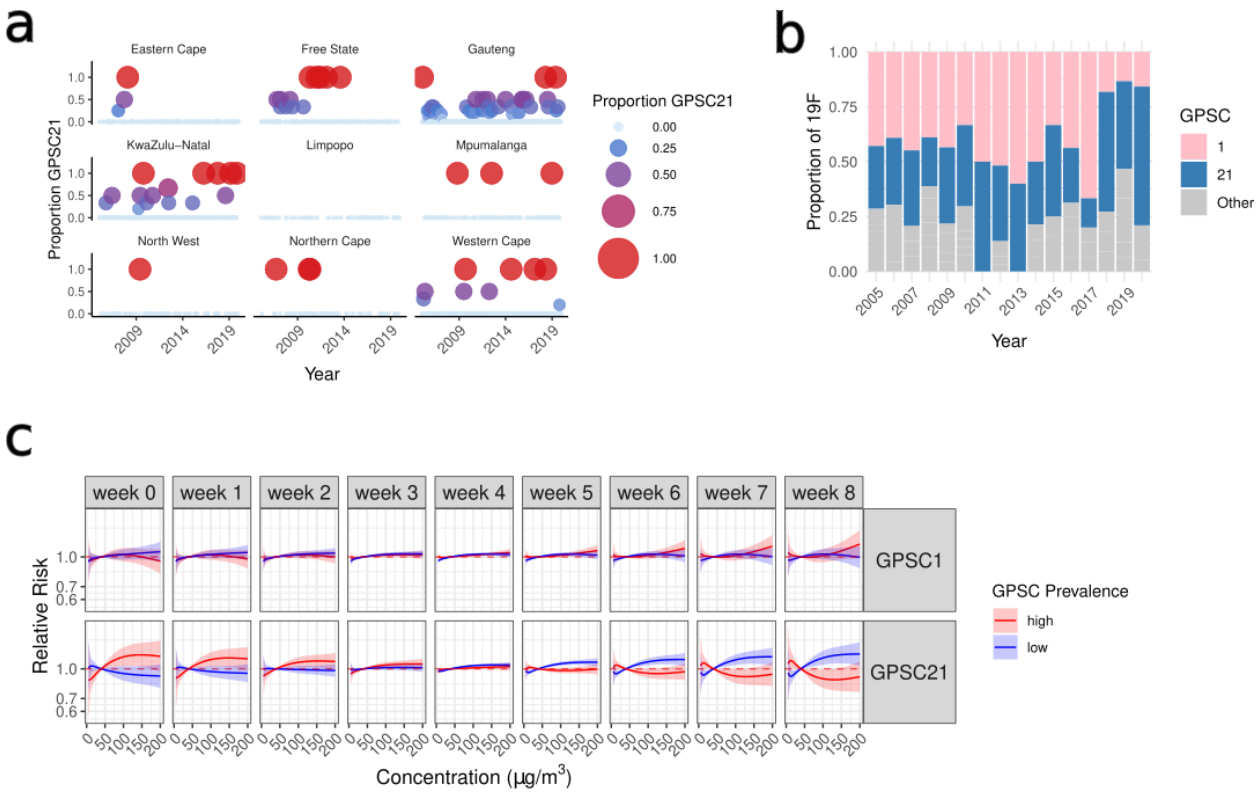

Figure 5. Incorporating GPSC21 by modifying the effect of $PM_{2.5}$ on IPD by the (a) proportion of sequenced which were GPSC21 in each week and province (b) The proportion of serotype 19F isolates which are found on each GPSC per year highlighting GPSC1 (pink) and GPSC21 (blue) (c) exposure-response curves for the relative risk across concentrations of PM2.5 for GPSC1 (top) and GPSC21 (bottom) including the curves for high prevalence (red) and low prevalence (blue) of each lineages.

Discussion

In this study, we applied spatio-temporal Bayesian hierarchical models to investigate the relationship between $PM_{2.5}$ exposure and the risk of IPD, while accounting for microbial diversity and key interventions (ART and PCV). In all presented models we accounted for interannual variability (replicated across provinces), seasonality, spatial autocorrelation, vaccination periods, and population density. We quantified the impact of ART and PCV on interannual variation in IPD, identified the effect of $PM_{2.5}$ with variable lag-times of the effects across pneumococcal serotypes, age groups, and disease types, and found the greatest increased risk in older adults. We examined whether genomic lineages modified the effects of $PM_{2.5}$ on IPD and found a higher short-term, acute (lag 0) effect of $PM_{2.5}$ when GPSC21-19F was highly prevalent (RR = 1.04; 95% CI 1.01–1.06).

We identified significant protective effects of ART and PCV on IPD although ARTs exacerbated impact on post-2020 NVTs was unexpected. It may reflect the increase in adult IPD cases in that period, and a waning effect of PCV (the major decreases in VT serotypes occurred at earlier time points[22]). Overall the impact of ART on opportunistic infections cannot be understated[17,23–25]. Funding for programs which have enabled ART coverage to exceed 70% in South Africa are being diminished and there are shifting funding priorities despite the many important secondary benefits of ART. Cuts to the U.S. President Emergency Plan for AIDS Relief (PEPFAR) in 2025 has been estimated to disrupt daily ART to at least 222,000 people and could lead to an increase of 30-64,000 new HIV infections by 2028[26]. There will likely increase in IPD incidence with the loss of funding for ART.

Our study confirmed a positive relationship between $PM_{2.5}$ and IPD with similar positive associations with PM_{10}, and SO_2. We focused the study on $PM_{2.5}$ due to sources often being anthropogenic and amenable to action[27–29]. The impact of underlying pneumococcal population diversity on the effect of air pollution has not previously been explored[30]. Notably, when we stratified the models by serotype we found an immediate effect from $PM_{2.5}$ on some specific serotypes including NVT serotype 8 and vaccine serotypes 4 and 23F but found a lagged effect for other serotypes. Different pneumococcal serotypes are known to vary in their invasive potential implying variability in disease mechanism[31]. The mechanism behind the effect of $PM_{2.5}$ on IPD is likely both on the host side (e.g. increased susceptibility to colonization[11], decreased mucociliary clearance[32]) and the bacterial side (e.g. increased transformation capacity[33] and biofilm formation[34])[13]. The immediate risk from $PM_{2.5}$ on NVTs and specific VTs may imply that there is a fitness advantage for these serotypes in high pollution areas. The lagged effect among most VTs suggests that vaccines provide partial protection against immediate $PM_{2.5}$-associated risk. The $PM_{2.5}$ effects persist across all age groups independently. The most acute effect (immediate risk from $PM_{2.5}$) is among those aged 15-64 years and the cumulative effect is greatest for those $\geq$65 years-old.

Importantly, the risk for children <5 years-old is non-negligible. These associations have been previously described for adults and children, with highest risk for older adults[28,35,36].

The pneumococcus is highly diverse and while the serotype data can summarize some of that diversity the immense recombination and dynamic population structure of the pneumococcus cannot be ignored. GPSCs have different carriage durations and demographic propensities[3,4]. It was not possible to directly stratify models including different GPSCs as an outcome, due to the high diversity (low-per-GPSC N), so we instead evaluated the interaction between environmental factors and GPSC prevalence. This allowed us to develop converging models adjusting for the underlying genomic population diversity.

GPSC21-19F is an emergent strain post-PCV in South Africa. The persistent serotype 19F (including PCV) is more frequently expressed by GPSC21 than the pre-PCV GPSC1-19F strain. Serotype 19F is especially prevalent among 5-18 year olds and behavioral differences may contribute to the acute effects of $PM_{2.5}$ on GPSC21-IPD[22]. Direct experimental evidence linking lineage-specific mechanistic data on $PM_{2.5}$-lineage data is currently lacking. Expanding genomic surveillance can support more mechanistic studies to investigate the specific biological interactions which may promote a $PM_{2.5}$ fitness advantage for particular lineages.

The concentrations of $PM_{2.5}$ and PM_{10} found in South Africa, especially the province of Gauteng, are extremely high. The strong positive association observed in this study is likely partially due to the high exposure concentration. The current threshold aim for $PM_{2.5}$ in South Africa is 40 $\mu g/m^3$ whereas the WHO threshold is 15 $\mu g/m^3$. We found in the year 2019 that 14.8 cases might have been averted if the South African threshold was met (24.3 cases by meeting the WHO threshold). More stringent air pollution goals could directly mitigate IPD.

There is an inextricable, bidirectional link between $PM_{2.5}$ and climate change in that particulate matter influences the Earth's radiative balance and atmospheric chemistry, while changes in temperature, precipitation, and land use affect the formation, composition, and distribution of PM[37,38]. Further, burning of fossil fuels both increases air pollution and global temperatures. There is biological plausibility for air pollution directly impacting host susceptibility, and promoting a fitness advantage for particular microbial traits. Our findings show that high levels of air pollution create conditions that favor particular serotypes and lineages. We provide microbial diversity-aware estimates of the risk of IPD driven by air pollution, demonstrating that the risk is not uniform across human populations with variable effects on at-risk groups and disease types. These interactions should be studied further across different regions with varied climatic conditions and demographic structures. Mechanistic experiments could enhance our understanding of the particular biology leading to these interactions between microbial diversity, host susceptibility, and environment.

The time period, high spatial resolution, and extensive metadata of our study is an asset for robust space-time ecological modelling studies. We believe that this novel approach for including genomic data could be leveraged for other infectious diseases to incorporate microbial genomic data into environmental epidemiological models.

Limitations

The reanalysis and observational datasets for $PM_{2.5}$ are imperfect. As we demonstrate in our comparison, air quality reanalysis models do not correlate well with the observational data across all sites. The available observational data from South African Air Quality Information Services (SAAQIS) across this time period is sparse and limits our ability to select an ideal reanalysis product. The correlation between CAMS and the observations was better than MERRA-2 and other bias corrected datasets. The absence of reanalysis models which are bias corrected to this region, and lack of observational data points for many less-urban regions, highlights the limitations of addressing the environmental impact of high-burden infectious diseases in low-income settings, with uncorrected environmental data[39,40]. To improve understanding of the environment's impact on high mortality infectious diseases efforts must be made to improve air pollution global reanalysis models[41]. The air quality standards we used as baseline values are the 24-hour standards set by NAAQS. These standards are higher than the health standards set by the WHO due to the high concentration of pollutants in some South African cities. As such, our relative risk estimates are calculated against threshold defined for the South African context but would be higher if compared to the WHO standards. We used weekly means rather than the 24-hour means for air quality metrics so these thresholds are lenient (e.g. if 1 day within a 7 day period has a 24 hour concentration exceeding the threshold but the other 6 days do not, the weekly mean concentration will likely be below the threshold).

Environmental reanalysis datasets should be expanded and refined across a wider range of geographies. This may require targeted investment in sustainable air quality monitoring networks, but the resulting evidence would elucidate air quality evolution and the impact on human health globally. Despite our inclusion of the entire GERMS-SA dataset across the time periods, sources of ascertainment bias include access to care, and correct processing and acquisition of specimens. Our data is from hospital location so that in cases of long-distance travel to hospitals cases may be attributed to a spatial grid different from the patient's home location. While audits were conducted to capture unreported cases these excluded serotype data. While our study included a large proportion of genomic sequences these did not encompass the entire epidemiological dataset. The isolates selected for genomic sequencing were randomly selected according to specific age groups in each sequencing set (as described in the methods). By incorporating the genomic lineages as a modifying proportional interaction effect, and normalizing this by the total

number of sequences in each space-time increment we try to limit these biases. We did not include carriage isolates in this analysis, so our results pertain only to IPD lineages and serotypes. We assume that the chosen lag window (8-weeks) for relevant covariates is sufficient to capture delayed effects and effects beyond this are assumed negligible.

Conclusion

In summary, environmental factors influence IPD risk in South Africa with humidity and particulate matter posing the highest disease risk. The relationships we found are delayed and non-linear, and vary given the serotypes or lineage circulating. While there was a strong association between $PM_{2.5}$ and IPD in all age groups independently, the highest risk is among those ≥65, and the most immediate risk is for those between 15 and 64. The lagged and variable effects indicate that the mechanism by which $PM_{2.5}$ increases IPD risk varies across groups (e.g., ages, serotypes, immunity). This study provides strong evidence that diverse bacterial populations are not uniformly affected by the environment. IPD patterns are influenced by increasing urban populations and climate change which may have important implications for disentangling vaccine escape, vaccine implementation strategies, and for hospital capacity preparations in changing environments. Critically, it also highlights the need to meet the policy thresholds for $PM_{2.5}$ to directly improve human health.

Materials and Methods

Data Processing

Pneumococcal Disease Datasets

Epidemiological Cases

The case counts of invasive pneumococcal disease (IPD) were provided by the National Institute for Communicable Diseases as part of national surveillance encompassing all nine provinces GERMS-SA. We included the entire dataset which comprised culture positive with viable isolates available for further analysis, culture positive with no viable isolates available, and culture negative with metadata results from PCR. Isolates were phenotypically characterised as has been previously described[22,42]. Finally, these included pneumococcal disease line list cases (N=59,017) between 2005-2023 collected from 531 hospitals across 52 districts in South Africa (Figure 1A-B). These excluded N=661 cases from the original dataset without hospital geolocation. Most cases were found in Gauteng with 24,303 cases. Northern Cape

reported the least number of cases, with 1,067 (Figure S1C). Disease types included 32,639 bacteremia (55.3%), 21,167 cases of meningitis (35.9%), and 5,211 other IPD cases (8.8%), such as empyema, septic arthritis, endophthalmitis, peritonitis, or pericarditis. These were classified given specimen type whereby meningitis cases were isolated from cerebrospinal fluid specimens, bacteremia from blood specimens, and other invasive disease classes were from other typically sterile sites. The median age was 32 with 13% of cases reported in children <1 year old. There was missing age data for 4.3% of cases and serotype information was missing for 37.8% of cases (Figure 1C). We included data from 2005-2023 in concordance with the environmental datasets for the initial baseline epidemiological model testing. As the genomic dataset ran until 2020, the year marking the beginning of the SARS-CoV-2 pandemic (with a substantial decrease in IPD counts), we restricted the models to the period 2005–2019 (N=52,437) so as to overlap with epidemiological cases and exclude this case reduction. In the overall environmental epidemiology models we included all data and for the models stratifications by outcome we included all reported data for each outcome. The outcomes we assessed included different age groups, disease types, and serotypes.

Table 1. Vaccine implementation year, schedule, and serotypes in South Africa. 2+1 indicates a 2+1 schedule whereby children receive the first two doses at 6 and 14 weeks of age and then a booster dose at 9 months.

Vaccine (introduction year in South Africa and schedule)	**Serotypes**
PCV7 (2009; 2+1)	4, 6B, 9V, 14, 18C, 19F, 23F
PCV13 (2011, 2+1)	4, 6B, 9V, 14, 18C, 19F, 23F, 1, 3, 5, 6A, 7F, 19A

Genomic Sequences

The genomes included in this study were sequenced as part of the Global Pneumococcal Sequencing project, GPS[43] which is a global genomic survey of *S. pneumoniae*. Isolates were collected by the National Institute for Communicable Disease in South Africa from 2005-2020 (n=4,912) from 48 of the 52 districts. We only include genomes through the year 2019 (N=4,350). Isolates were randomly sampled using a previously described protocol whereby approximately 300 invasive-disease isolates from each year (2005–2014) were selected with a specific target age breakdown (50% from <3 year olds, 25% from 3–5 year olds, 25% from >5 year olds). In the second phase of sequencing, 100 disease isolates from 2005 to 2010 from children <5 years old were additionally, randomly sequenced. The sampling for 2015-2020 was similar however the sampling strategy was not adjusted for the reduction in IPD cases following PCV (thus included more >5 year-olds)[6,22,44]. Most sequences are from the districts with the highest population

size (Figure S11). Assembly quality control parameters included a minimum average sequencing depth of 20× and an assembly length of 1.9–2.3 Mb. Sequences with more than 220 heterozygous SNP sites were excluded. The genomes were assembled using Spades assembly software[45]and the gene annotations were made using Prokka[46] PopPunk[47]was used to assign each genome to a Global Pneumococcal Sequence Cluster (GPSC). We included each GPSC with an N>90 and included GPSC8 due to its public health relevance[22]. We included the proportion as the number of each GPSC over the total number of sequences per week per province (Figure S12).

Sociodemographic Data

The sociodemographic data was extracted from multiple sources. The population sizes per year per district were sourced from the 2023 Statistics South Africa projections[48,49] From the same report we included per province ART coverage, HIV prevalence, life expectancy, growth rate, % children (<15) and the % elderly over time (60+). We also included the aging index and dependency ratio. These are defined by Statistics South Africa as the number over age 65 per 100 youths and the number <15 years old combined with the number 60+ over the number of working age population (15-64) respectively.[49] The pneumococcal conjugate vaccine has a 2+1 schedule in South Africa and was first introduced with PCV7 in 2009 and PCV13 in 2011 (Table 1). PCV coverage for 1, 2, and 3 doses were included from the WHO Immunization Data report.[50] Three doses is defined as the percentage of surviving infants who receive the 3rd dose of pneumococcal conjugate vaccines. We included the official coverage estimates as well as the WHO/UNICEF Estimates of National Coverage (WUENIC)[51] (Figure S1D). The official estimates initially are based on surveys and other administrative messages, which, due to biases and inaccuracies, are adjusted by the national authorities to provide estimates of the most likely coverage.

Environmental Datasets

Meteorological Data

The environmental datasets spanned 2005-2023 inclusive. We included temperature, precipitation, relative, and absolute humidity metrics from the global reanalysis dataset ERA5-Land[52] aggregated for weekly time points at the district level. This is gridded data with 0.1° degree resolution (accounting for the mean latitude of South Africa at -29° so each grid square is approximately 108km^2). This resulted in a mean of 218 grid cells per district ranging from 15 in Johannesburg, Gauteng to 1180 in Namakwa, Northern Cape. Precipitation was initially retrieved as hourly average meters (m) per second precipitation loss rate (prlr). We converted it to daily millimeters (mm).

We calculated the Standardised Precipitation Index (SPI) and Standardised Precipitation and Evapotranspiration Index (SPEI) at 1-month, 3-month, 6-month, and 12-month time periods using the above mentioned precipitation and temperature data from ERA5-Land. The packages used for this calculation were SPEI[53], ClimProjDiags[54], and CSTools[55].

Absolute humidity (absh) is defined as the mass of water vapour in a volume of air and the units are g/m^3. Relative humidity (hurs) is the water vapour present in the air relative to the saturated air and defined as the partial pressure of water vapour as a percentage of the saturation pressure (%)(Figure 1D).

Air Pollution Data

The air pollution datasets spanned 2005-2023 inclusive. To determine which estimates of air pollution to include we compared air pollution estimates obtained from Copernicus Atmosphere Monitoring Service (CAMS) global reanalysis dataset, and NASA's MERRA-2[56,57] We performed an assessment of the correlation between the gridded air quality reanalysis models against observational data made available by the South African Air Quality Information Services (SAAQIS)[58]. $PM_{2.5}$ and PM_{10} are not directly included in MERRA-2, so they were computed using methods detailed in the MERRA-2 documentation[59](Figure 1D)(Table 2).

Table 2. Description of air quality reanalysis products. Calculations to adjust each air-quality reanalysis product to $\mu g/m^3$ detailing the mass used for $PM_{2.5}$, PM_{10}, O_3, and SO_2.

Variable	**CAMS (50km^2)**		**MERRA (80km^2)**		**MERRA_BiasFix (80km^2)**	
Adjustment	Initial	Adjust# to $\mu g/m^3$	Initial	Adjust# to $\mu g/m^3$	Initial	Adjust to $\mu g/m^3$
$PM_{2.5}$	kg m $^{-3}$	$*10^9$	kg m $^{-3}$	$*10^9$	kg m $^{-3}$	$*10^9$
PM_{10}	kg m $^{-3}$	$*10^9$	kg m $^{-3}$	$*10^9$	-	-
O_3	kg kg $^{-1}$	$*2.14*10^9$	kg kg $^{-1}$	$*2.14*10^9$	-	-
SO_2	kg kg $^{-1}$	$*2.63*10^9$	kg kg $^{-1}$	$*2.63*10^9$	-	-
# air density volume included as 1.23, O_3 density as 2.14 and SO_2 as 2.63						

We included the MERRA-2 product from NASA for the entire time period and also included the bias corrected MERRA-2 product from 2020-2023[60] The CAMS global datasets are at approximately an

$80km^2$ resolution while the MERRA-2 datasets are at a $50km^2$ spatial resolution. We also included a gapless 1km gridded machine learning model from Wei *et al.,* 2023[61]

We extracted observational data from air quality monitoring stations provided by SAAQIS in South Africa. We calculated a mean absolute error, root mean square error, and Spearman's correlation between the reanalysis products and the mean across observations in the district of Gert Sibande, Mpumalanga, Gauteng province, and KwaZulu-Natal province (Figure S15A). These were the locations with the most consistent observations across the time period available in SAAQIS. We performed a formal comparison of the datasets against observational data and found that CAMS was the best performing reanalysis model (Figure S15B). CAMS performed best despite the poor performance in the Gert Sibande region where none of the products performed well, either due to the observations being flawed, or a bias across all reanalysis products impacting this region. We included the air quality reanalysis data from CAMS in our study (Table 3).

The R package startR was used to retrieve and reformat all multidimensional environmental datasets[62].

Table 3. Correlations between observational station data and each air quality analysis product. Relationship between reanalysis products from Copernicus (CAMS), MERRA-2 from NASA, MERRA-2 with a bias adjustment and a machine learning gapless product at 1km grid squares from Wei et al. 2023. We include a spearman's correlation, mean absolute error ($\mu g/m^3$) and the root mean squared error ($\mu g/m^3$).

	Gauteng				KwaZulu-Natal				Gert Sibande, Mpumalanga			
Statistic	CAMS	MERRA-2	MERRA-2 -adjusted	ML Gapless 1km	CAMS	MERRA-2	MERRA-2 -adjusted	ML Gapless 1km	CAMS	MERRA-2	MERRA-2 -adjusted	ML Gapless 1km
Correlation	0.88	0.49	0.58	0.82	0.58	0.52	0.17	0.08	0.08	0.30	-0.21	0.27
MAE	43.06	74.79	62.05	55.52	32.49	52.98	50.74	48.15	74.55	92.95	83.27	83.40
RMSE	58.46	81.02	67.47	61.40	42.34	59.69	57.50	55.69	105.45	119.56	113.56	111.66

Epidemiological Model Specification

Bayesian hierarchical mixed models using integrated nested laplace approximation

Statistical Analysis

All statistical analysis and visualizations were conducted using R version 4.4.2[63] with key packages including R-INLA[64], dlnm[65], dplyr[66], sf[67], and vegan[68].

Base Model Evaluation

We modeled weekly pneumococcal disease counts in South Africa between 2005-2023 (19 years) by specifying a Bayesian hierarchical mixed model using pneumococcal disease case counts at the district (N=52) or province (N=9) level as the response variable. To account for overdispersion, we assumed that case counts follow a negative binomial distribution such that:

$$y_{s,t} \mid \mu_{s,t} \sim \text{NegBin}(\mu_{s,t}, \kappa)$$
$$\log(\mu_{s,t}) = \log(p_{s,a(t)}) + \log(\rho_{s,t})$$

where $\mu_{s,t}$ is the expected number of cases and κ is the overdispersion parameter. The distribution mean, $\mu_{s,t}$, is defined as the product of the population $p_{s,a(t)}$ (per 100,000 people) in a given spatial unit, s, (district or province) and year a(t), and the disease incidence $\rho_{s,t}$.

We tested increasingly complex baseline models ultimately selecting M12 as our baseline model (Table 4)(Table S2). We evaluated the model performance defining the response variable both at district and province level for weekly and monthly case counts and across various random and fixed effect combinations. We proceeded with the weekly models due to their relevance for respiratory disease effects (Table S2).

Baseline model, M12, included an intercept α as well as spatiotemporal random effects accounting for unknown and unmeasured monthly effects (calendar months January (1) to December (12)), $\delta_{m(t)}$, and interannual effect replicated by province (n=9 provinces) $\gamma_{a(t,pr)}$. This model also included spatial effects with a Besag-York-Mollié (BYM-2)[69] including both structured (u_s) and unstructured effects (v_s).

We included a single categorical covariate with three-phases to account for decreasing case counts after the implementation of the pneumococcal conjugate vaccine (PCV): pre-PCV (years before 2009), PCV7 (2009-2010 inclusive), and PCV13 (2011 onwards). This accounted for a non-linear effect from vaccination. We also included district level human population density (*pd*) as a covariate due to its particular relevance as a confounding variable with particulate matter. In the province level model $s = pr$.

$$\log(\rho_{t,s}) = \alpha + \delta_{m(t)} + \gamma_{a(t,pr)} + u_s + v_s + pcv_{periods} + pd_{s,a(t)}$$

We evaluated models with different random effect formulations and compared them to the intercept only model. We found an increasing goodness-of-fit with increasing complexity (Table S2). We selected our baseline model due to the balance of complex random effects and goodness-of-fit.

Table 4. Model formulations with increasing complexity to determine baseline model at a weekly temporal resolution and including years 2005-2019. WAIC = Watanabe-Akaike Information Criterion, MAE = Mean Absolute Error, CPO = log10 conditional predictive ordinate, and R^2 likelihood ratio compared to the intercept model.

Model	Formula	WAIC	MAE	CPO	R^2
M0	$\log_{(\rho t,s)} = \alpha$	99687.81	1.18	1.23	0
M1	$\log_{(\rho t,s)} = \alpha + \delta_{m(t)} + \gamma_{a(t)}$	97301.75	1.14	1.2	0.06
M2	$\log_{(\rho t,s)} = \alpha + \delta_{m(t)} + u_s + v_s$	84082.5	0.9	1.04	0.32
M3	$\log_{(\rho t,s)} = \alpha + \gamma_{a(t)} + u_s + v_s$	82439.36	0.8	1.02	0.35
M4	$\log_{(\rho t,s)} = \alpha + \delta_{m(t)} + \gamma_{a(t)} + u_s + v_s$	80883	0.76	1	0.37
M5	$\log_{(\rho t,s)} = \alpha + \delta_{m(t)} + \gamma_{a(t)} + u_s + v_s + pcv_{2009vaccine}$	80882.95	0.76	1	0.37
M6	$\log_{(\rho t,s)} = \alpha + \delta_{m(t)} + \gamma_{a(t)} + u_s + v_s + pcv_{periods}$	80882.89	0.76	1	0.37
M7	$\log_{(\rho t,s)} = \alpha + \delta_{m(t)} + \gamma_{a(t)} + u_s + v_s + pcv_{periods} + covid$	80883.01	0.76	1	0.37
M8	$\log_{(\rho t,s)} = \alpha + \delta_{m(t,pr)} + \gamma_{a(t)} + u_s + v_s + pcv_{periods}$	80794.05	0.75	1	0.38
M9	$\log_{(\rho t,s)} = \alpha + \delta_{m(t)} + \gamma_{a(t,pr)} + u_s + v_s$	78943.71	0.7	0.97	0.41
M10	$\log_{(\rho t,s)} = \alpha + \delta_{m(t)} + \gamma_{a(t,pr)} + u_s + v_s + pcv_{2009vaccine}$	78942.09	0.7	0.97	0.41
M11	$\log_{(\rho t,s)} = \alpha + \delta_{m(t)} + \gamma_{a(t,pr)} + u_s + v_s + pcv_{periods}$	78939.5	0.7	0.97	0.41
M12	$\log_{(\rho t,s)} = \alpha + \delta_{m(t)} + \gamma_{a(t,pr)} + u_s + v_s + pcv_{periods} + pd_{s,a(t)}$	78766.03	0.7	0.97	0.41
M13	$\log_{(\rho t,s)} = \alpha + \delta_{m(t)} + \gamma_{a(t,pr)} + u_s + v_s + pcv_{national}$	78940.44	0.70	0.97	0.41
M14	$\log_{(\rho t,s)} = \alpha + \delta_{m(t)} + \gamma_{a(t,pr)} + u_s + v_s + art_{national}$	78939.15	0.70	0.97	0.41
M15	$\log_{(\rho t,s)} = \alpha + \delta_{m(t)} + \gamma_{a(t,pr)} + u_s + v_s + art_{province}$	78938.82	0.70	0.97	0.41
M16	$\log_{(\rho t,s)} = \alpha + \delta_{m(t)} + \gamma_{a(t,pr)} + u_s + v_s + art_{province} + pcv_{periods}$	78938.55	0.70	0.97	0.41

Impact of antiretroviral therapy and pneumococcal vaccine on interannual variation

To assess the impact of the uptake of ART (treatment for HIV) across South Africa compared to the implementation of the vaccine we evaluated models with them included independently and together. We tested different models including the ART and PCV covariates with different formulations both together and independently. These included: i) PCV as a categorical variable with 2 categories pre-PCV7 (2005-2008), and post-PCV (2009-2023) (M10); ii) PCV as a single categorical covariate with 3 categories for three vaccine periods (pre–PCV7, PCV7, and PCV13) (M11); iii) PCV coverage with one value (% coverage) nationally per year (M13); iv) ART coverage with one value (% coverage) nationally per year (M14); v) ART coverage with one value (% coverage), per province, per year (M15); vi) and a

model including ART % coverage per province/per year and a categorical covariate covering the 3 vaccination periods (M16) (Table 4). These were all run with and without a replication of the interannual effect across provinces (Table S4). We also ran different models for different types of IPD to assess the impact of these perturbations across different groups. These included children ≤5 years old, adults between 15 and 64 years old, adults 65 and above, IPD from only NVTs and from only VTs. We also ran these models including disease from 2005-2019 (excluding the post-COVID-19 period), and from 2005-2023 (Table S4).

Distributed lag non-linear models to estimate disease risk with environmental factors

Overall Models

To account for lag times from the exposure to disease outcome we specified distributed lag non-linear models (DLNMs) to estimate risk of pneumococcal disease incidence for different environmental factors and lag times[31,354]. To explore the data, we correlated lag times from exposure to disease outcome using a Spearman correlation (Figure S16). We included cases from 2005-2019 to match our genomic sequence models. We ran weekly models at both province (adm1) and district (adm2) level. We evaluated 6, 8, and 12 week lag times across the variables and administrative regions finding that 8 weeks of lags provided the best model fit. We generated the DLNM crossbasis with 8 weeks of lag with 2 degrees of freedom for the lag knots except for relative humidity for which we included 3 degrees of freedom to capture the complex non-linearity. For each variable we included equally distributed knots at the 33rd and 67th quantile except for precipitation for which we included 2 equally spaced knots on a log scale of the precipitation distribution. We centered and scaled each environmental variable. We evaluated the Pearson correlation between the environmental variables. We also tested models including the province of Gauteng alone due to it being the most populous, highest incidence of IPD, and highest maximum concentrations of $PM_{2.5}$, PM_{10}, and SO_2 (Figure S4)(Figure S5).

For the best models (relative humidity, absolute humidity, $PM_{2.5}$ and PM_{10}) with $R^2>0.6$ we evaluated the inclusion of additional environmental covariates (maximum temperature, $PM_{2.5}$, PM_{10}, SO_2) as linear terms. We also tested the addition of relative humidity and absolute humidity as non-linear effects using random walk 2 models with 5 cuts. Because the model uses a log-link, additive terms in the linear predictor correspond to multiplicative effects at the response level (i.e. disease incidence rates). We estimated the exposure-lag-response association using the crosspred function. For air quality metrics we centered at the 24-hour National Ambient Air Quality Standards (NAAQS) for South Africa ($PM_{2.5} = 40$, $PM_{10} = 75$, and $SO_2 = 125$ μg/m^3). We utilized the 24 hour exceedance threshold rather than the annual as

we are examining an acute effect (IPD) (Table 5). For the rest we centered this estimation at the median of the variable distribution (Table S5).

Table 5. Air quality standards. For the years 2016-2029 as set by the Environmental Management Air Quality Act 39 in South Africa and the standards set by the WHO in 2021.

Air Quality Component	SA 24-hour	WHO 24-hour
$PM_{2.5}$	40	15
PM_{10}	75	45
SO_2	125	40
**Standards set by the South African environmental management air quality act for 2016-2029*[21]		

Stratification by age, disease, and vaccine status

To identify differences in risk across different stratification groups we changed our outcome from IPD overall to disease by age group, disease outcome, vaccine status (whether cases were serotypes which were included in the PCV7 vaccine, PCV13 vaccine, or were non-vaccine type), and by specific serotype. For these we included the M12 baseline model (Table 4), but we changed the outcome to disease incidence among each of these groups per 100,000 population. The models tested included i) age $\leq$5 years old, ii) age 15-64 years old, iii) age $\geq$65, iv) bacteremia, v) meningitis, vii) other invasive disease types. We also included serotype stratifications by vii) NVT, viii) PCV7, and ix) PCV13. Models 11-28 included stratification by IPD from particular serotypes (Table S1). We initially included ages 6 to 14 and found no effect across any lag-time or cumulatively with large credible intervals so exclude this group in the main results.

To estimate the relative risk ratio between each subgroup (serotype, disease type, age group) compared to the overall population we crosscut the relative risk from the exposure-response curve at 50 $\mu g/m^3$ concentration for each. We then included the relative risk for each subgroup in turn as the numerator compared to the relative risk from the overall model to estimate the increased (>1) or decreased (<1) risk of IPD, in each subgroup, from $PM_{2.5}$ compared to the risk for pneumococcal disease overall (Figure 4A).

Distributed lag non-linear models interacting with lineage prevalence

To assess the impact of underlying lineage diversity on the environmental effects we utilized the same baseline model (M12: seasonal, spatial, and province-replicated-interannual random effects, together with vaccination periods and population density)(Table 4). We then included an interaction between the exposure crossbasis and a modifying variable representing GPSC-specific proportions per total number

sequenced each week. We only included a province level spatial effect for these interaction models. To ensure model interpretability, we scaled and centered this variable (Table S6). We only included space-time units in which each modifying variable is non-zero.

We evaluated the exposure–response relationship at three representative values of the effect modifier, corresponding to the 0.1, 0.5, and 0.8 quantiles of the centered distribution of the modifying variable (excluding zeros). For each quantile value z, we derived level-specific coefficient vectors and corresponding variance–covariance matrices that reflect the combined main and interaction effects.

To account for uncertainty in the interaction effects across different levels of the modifying variable, we computed level-specific coefficient vectors and their corresponding variance-covariance matrices. For a given value z of the centered interaction variable, the coefficient vector was calculated as:

$$\beta_z = \beta_0 + z \bullet \beta_{int}$$

and the associated variance-covariance matrix as:

$$V_z = V_0 + z^2 \bullet V_{int} + 2z \bullet V_{cross}$$

where β_0 and V_0 are the estimated coefficients and variance-covariance matrix for the main (non-interaction) effect, β_{int} and V_{int} are those for the interaction term, and V_{cross} represents the covariance between main and interaction coefficients. These were then passed to the crosspred() function of the dlnm package to generate estimates and confidence intervals at each value of z (Table S6).

We performed a sensitivity analysis whereby for GPSC pairs, which share the same dominant serotypes, we split the dataset into the pre- and post-vaccine period to ensure the effect was not just driven by serotype replacement in the time series but by some underlying evolutionary process (Figure S14).

To estimate goodness-of-fit for each model we calculated the Watanabe-Akaike information criterion (WAIC)[71]and the mean absolute error (MAE). We also calculated the sum of the log10 conditional predictive ordinate (CPO) for each observation to evaluate the model's out-of-sample predictive performance. We additionally computed a pseudo-R^2, defined as the proportionate reduction in deviance compared to an intercept-only model[72] This measured the ability of each model to account for variation in disease incidence (0-1) whereby a model with $R^2 = 1$ would indicate a model accounting for all of the variation in the data. This helped evaluate the improvement in model fit achieved by including additional covariates.

Data Availability

All code is available in the GitHub repository (https://github.com/sophbel/env_sa_manuscript).

Competing Interests

The authors report no conflicts of interest.

Authors' contributions

SB and RL conceptualized the study. SB, CL, JK, DL, CC, SM, MdP, AvG contributed to data curation and harmonization. SB, CL, JK, DL, CC, SM, MdP, GM, AvG, RL were responsible for study design. CL, JK, SM, MdP, AvG provided expert knowledge and insight from the study location. SB, GM, RL contributed to methodological development. SB did the formal analysis. SB wrote the draft manuscript. All authors edited, reviewed, and approved the final manuscript.

Acknowledgements

This work was supported in part by a Fogarty International Center Global Infectious Disease research training grant, National Institutes of Health, to the University of Pittsburgh and National Institute for Communicable Diseases (D43TW011255). Genome sequencing was funded by the Gates Foundation (grant code OPP1189062) and Wellcome Trust (Grant reference 206194) through the Global Pneumococcal Sequencing project, and the SEQAFRICA project which was funded by the Department of Health and Social Care's Fleming Fund using UK aid. The views expressed in this publication are those of the authors and not necessarily those of the UK Department of Health and Social Care or its Management Agent, Mott MacDonald.

S.B. is supported by Schmidt Science Fellows, in partnership with the Rhodes Trust. C.C. and D.L. acknowledges their AI4S fellowship within the "Generación D" initiative by Red.es, Ministerio para la Transformación Digital y de la Función Pública, for talent attraction (C005/24-EDCV1), funded by NextGenerationEU through PRTR. R.L. acknowledges support from the Wellcome Trust (IDExtremes 226069/Z/22/Z), the European Union's Horizon Europe research and innovation program (IDAlert, grant agreement 101057554) and a Royal Society Dorothy Hodgkin Fellowship.

In addition, we would like to thank the CRDM surveillance team at the NICD for assisting with retrieval of isolates, phenotypic characterization, managing and cleaning of the GERMS database and the

Sequencing Core Facility at the NICD for the generation of additional genomic data. We would also like to thank the Atmospheric Composition group at the Barcelona Supercomputing Center for assistance with atmospheric dataset acquisition and processing. We would also like to thank Professor Rebecca Garland at the University of Pretoria, South Africa, for useful discussions regarding the performance of global reanalysis models in South Africa.

For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

1. Weiser, J. N., Ferreira, D. M. & Paton, J. C. Streptococcus pneumoniae: transmission, colonization and invasion. *Nat. Rev. Microbiol.* **16**, 355–367 (2018).
2. Wahl, B. *et al.* Burden of Streptococcus pneumoniae and Haemophilus influenzae type b disease in children in the era of conjugate vaccines: global, regional, and national estimates for 2000-15. *Lancet Glob. Health* **6**, e744–e757 (2018).
3. Gladstone, R. A. *et al.* International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. *EBioMedicine* **43**, 338–346 (2019).
4. Lo, S. W. *et al.* Pneumococcal lineages associated with serotype replacement and antibiotic resistance in childhood invasive pneumococcal disease in the post-PCV13 era: an international whole-genome sequencing study. *Lancet Infect. Dis.* **19**, 759–769 (2019).
5. Lo, S. W. *et al.* Emergence of a multidrug-resistant and virulent Streptococcus pneumoniae lineage mediates serotype replacement after PCV13: an international whole-genome sequencing study. *Lancet Microbe* **3**, e735–e743 (2022).
6. Belman, S. *et al.* Geographic migration and vaccine-induced fitness changes of Streptococcus pneumoniae. 2023.01.18.524577 Preprint at https://doi.org/10.1101/2023.01.18.524577 (2023).
7. German, E. L. *et al.* Protective effect of PCV vaccine against experimental pneumococcal challenge in adults is primarily mediated by controlling colonisation density. *Vaccine* **37**, 3953–3956 (2019).
8. Malik, T. M. *et al.* A dynamic transmission model for assessing the impact of pneumococcal vaccination in the United States. *PLOS One* **20**, e0305892 (2025).
9. Oidtman, R. J. *et al.* Modelling the Epidemiological Impact of Different Adult Pneumococcal Vaccination Strategies in the United Kingdom. *Infect. Dis. Ther.* **14**, 587–602 (2025).
10. Russell, F. M., Mulholland, K. & Greenslade, L. Mass doses of fractional pneumococcal vaccine in humanitarian settings. *Lancet Infect. Dis.* **25**, 599–601 (2025).
11. Cheliotis, K. S. *et al.* Influence of sex, season and environmental air quality on experimental human pneumococcal carriage acquisition: a retrospective cohort analysis. *ERJ Open Res.* **8**, (2022).

12. Numminen, E. *et al.* Climate induces seasonality in pneumococcal transmission. *Sci. Rep.* **5**, 11344 (2015).
13. Beentjes, D., Shears, R. K., French, N., Neill, D. R. & Kadioglu, A. Mechanistic Insights into the Impact of Air Pollution on Pneumococcal Pathogenesis and Transmission. *Am. J. Respir. Crit. Care Med.* **206**, 1070–1080 (2022).
14. Rentschler, J. & Leonova, N. Global air pollution exposure and poverty. *Nat. Commun.* **14**, 4432 (2023).
15. Dube, F. S. *et al.* Longitudinal characterization of nasopharyngeal colonization with Streptococcus pneumoniae in a South African birth cohort post 13-valent pneumococcal conjugate vaccine implementation. *Sci. Rep.* **8**, 12497 (2018).
16. Downs, S. L. *et al.* Streptococcus pneumoniae and other bacterial nasopharyngeal colonization seven years post-introduction of 13-valent pneumococcal conjugate vaccine in South African children. *Int. J. Infect. Dis.* **134**, 45–52 (2023).
17. Meiring, S. *et al.* HIV Infection and the Epidemiology of Invasive Pneumococcal Disease (IPD) in South African Adults and Older Children Prior to the Introduction of a Pneumococcal Conjugate Vaccine (PCV). *PLOS ONE* **11**, e0149104 (2016).
18. Heinsbroek, E. *et al.* Persisting high prevalence of pneumococcal carriage among HIV-infected adults receiving antiretroviral therapy in Malawi: a cohort study. *AIDS Lond. Engl.* **29**, 1837–1844 (2015).
19. Bertran, M. *et al.* Invasive pneumococcal disease 3 years after introduction of a reduced 1 + 1 infant 13-valent pneumococcal conjugate vaccine immunisation schedule in England: a prospective national observational surveillance study. *Lancet Infect. Dis.* **24**, 546–556 (2024).
20. Bennett, J. C. *et al.* Global impact of ten-valent and 13-valent pneumococcal conjugate vaccines on invasive pneumococcal disease in all ages (the PSERENADE project): a global surveillance analysis. *Lancet Infect. Dis.* **25**, 457–470 (2025).
21. National Environmental Management: Air Quality Act: National ambient air quality standard for particulate matter with aerodynamic diameter less than 2.5 micron metres (PM 2.5) | South African

Government. https://www.gov.za/documents/national-environmental-management-air-quality-act-national-ambient-air-quality-standard-0.

22. Lekhuleni, C. *et al.* Impact of pneumococcal conjugate vaccines on invasive pneumococcal disease-causing lineages among South African children. *Nat. Commun.* **15**, 8401 (2024).
23. B-Lajoie, M.-R. *et al.* Incidence and Prevalence of Opportunistic and Other Infections and the Impact of Antiretroviral Therapy Among HIV-infected Children in Low- and Middle-income Countries: A Systematic Review and Meta-analysis. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* **62**, 1586–1594 (2016).
24. Low, A. *et al.* Incidence of Opportunistic Infections and the Impact of Antiretroviral Therapy Among HIV-Infected Adults in Low- and Middle-Income Countries: A Systematic Review and Meta-analysis. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* **62**, 1595–1603 (2016).
25. Akingbola, A., Adegbesan, A., Mariaria, P., Isaiah, O. & Adeyemi, E. A pause that hurts: the global impact of halting PEPFAR funding for HIV/AIDS programs. *Infect. Dis.* **57**, 378–384 (2025).
26. Comprehensive update on HIV programmes in South Africa | UNAIDS. https://www.unaids.org/en/resources/presscentre/featurestories/2025/february/20250225_south-africa-fs.
27. Bhattarai, H., Tai, A. P. K., Val Martin, M. & Yung, D. H. Y. Responses of fine particulate matter (PM2.5) air quality to future climate, land use, and emission changes: Insights from modeling across shared socioeconomic pathways. *Sci. Total Environ.* **948**, 174611 (2024).
28. Yang, T. *et al.* Epidemiological characteristics and meteorological factors of acute respiratory infections (ARIs) in hospitalized children in eastern Guangdong, China. *Sci. Rep.* **14**, 25518 (2024).
29. Jusot, J.-F. *et al.* Airborne dust and high temperatures are risk factors for invasive bacterial disease. *J. Allergy Clin. Immunol.* **139**, 977-986.e2 (2017).
30. Ma, J. *et al.* Fine particulate matter manipulates immune response to exacerbate microbial pathogenesis in the respiratory tract. *Eur. Respir. Rev.* **33**, 230259 (2024).

31. Darkwah, S., Somda, N. S., Mahazu, S. & Donkor, E. S. Pneumococcal serotypes and their association with death risk in invasive pneumococcal disease: a systematic review and meta-analysis. *Front. Med.* **12**, (2025).

32. Carvalho-Oliveira, R. *et al.* Chemical composition modulates the adverse effects of particles on the mucociliary epithelium. *Clinics* **70**, 706–713 (2015).

33. de Oliveira, J. D. *et al.* Influence of environmental dust in transformation capacity of Streptococcus pneumoniae. *Genet. Mol. Biol.* **44**, e20200028.

34. Yadav, M. K., Go, Y. Y., Chae, S.-W., Park, M. K. & Song, J.-J. Asian Sand Dust Particles Increased Pneumococcal Biofilm Formation in vitro and Colonization in Human Middle Ear Epithelial Cells and Rat Middle Ear Mucosa. *Front. Genet.* **11**, (2020).

35. Zhou, Z., Fang, C., Li, J., Zhou, M. & Chen, X. Ambient NO(2) is associated with Streptococcus pneumoniae-induced pneumonia in children and increases the minimum inhibitory concentration of penicillin. *Int. J. Biometeorol.* **66**, 87–96 (2022).

36. Dherani, M. K. *et al.* Association between household air pollution and nasopharyngeal pneumococcal carriage in Malawian infants (MSCAPE): a nested, prospective, observational study. *Lancet Glob. Health* **10**, e246–e256 (2022).

37. Bhattarai, H., Tai, A. P. K., Val Martin, M. & Yung, D. H. Y. Responses of fine particulate matter (PM2.5) air quality to future climate, land use, and emission changes: Insights from modeling across shared socioeconomic pathways. *Sci. Total Environ.* **948**, 174611 (2024).

38. Short-lived Climate Forcers. in *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (ed. Intergovernmental Panel on Climate Change (IPCC)) 817–922 (Cambridge University Press, Cambridge, 2023). doi:10.1017/9781009157896.008.

39. Garland, R. M., Naidoo, M., Sibiya, B. & Oosthuizen, R. Air quality indicators from the Environmental Performance Index: potential use and limitations in South Africa. *Clean Air J.* **27**, 33–33 (2017).

40. Garland, R., Wernecke, B., Feig, G. & Langerman, K. The new WHO Global Air Quality Guidelines: What do they mean for South Africa? *Clean Air J.* **31**, (2021).

41. Naghavi, M. *et al.* Global burden of 292 causes of death in 204 countries and territories and 660 subnational locations, 1990–2023: a systematic analysis for the Global Burden of Disease Study 2023. *The Lancet* **406**, 1811–1872 (2025).

42. Ndlangisa, K. M. *et al.* A Streptococcus pneumoniae lineage usually associated with pneumococcal conjugate vaccine (PCV) serotypes is the most common cause of serotype 35B invasive disease in South Africa, following routine use of PCV. *Microb. Genomics* **8**, 000746 (2022).

43. GPS Consortium. Global Pneumococcal Sequencing collection. *Microb. Genomics* https://www.microbiologyresearch.org/content/global-pneumococcal-sequencing-collection (2022).

44. Lekhuleni, C. N. *et al.* Impact of Pneumococcal Conjugate Vaccines on Invasive Pneumococcal Disease-Causing Lineages in South African Children, 2005–20: A National Genomic Epidemiology Study. SSRN Scholarly Paper at https://doi.org/10.2139/ssrn.4652526 (2023).

45. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

46. Prokka: rapid prokaryotic genome annotation | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/30/14/2068/2390517.

47. Lees, J. A. *et al.* Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* **29**, 304–316 (2019).

48. WorldPop, open data for spatial demography | Scientific Data. https://www.nature.com/articles/sdata20174.

49. Africa, S. S. 2023 | Statistics South Africa. https://www.statssa.gov.za/?m=2023 (2023).

50. WHO Immunization Data portal - Detail Page. *Immunization Data* https://immunizationdata.who.int/global/wiise-detail-page.

51. WHO/UNICEF estimates of national immunization coverage. https://www.who.int/teams/immunization-vaccines-and-biologicals/immunization-analysis-and-insig

hts/global-monitoring/immunization-coverage/who-unicef-estimates-of-national-immunization-coverage.

52. Muñoz-Sabater, J. *et al.* ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data* **13**, 4349–4383 (2021).
53. Beguería, S. & Vicente-Serrano, S. M. SPEI: Calculation of the Standardized Precipitation-Evapotranspiration Index. (2023).
54. BSC-CNS *et al.* ClimProjDiags: Set of Tools to Compute Various Climate Indices. (2025).
55. Perez-Zanon, N. *et al.* CSTools: Assessing Skill of Climate Forecasts on Seasonal-to-Decadal Timescales. (2025).
56. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) in: Journal of Climate Volume 30 Issue 14 (2017). https://journals.ametsoc.org/view/journals/clim/30/14/jcli-d-16-0758.1.xml.
57. CAMS Air Quality Data – Quality Assured | Copernicus. https://atmosphere.copernicus.eu/cams-air-quality-data-quality-assured.
58. Saaqis. https://saaqis.environment.gov.za/.
59. MERRA-2. https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/.
60. Gupta, P. MERRA2_CNN_HAQAST bias corrected global hourly surface total PM2.5 mass concentration, V1. NASA Goddard Earth Sciences Data and Information Services Center https://doi.org/10.5067/OCKK5HCFW5N3 (2023).
61. Wei, J. *et al.* First close insight into global daily gapless 1 km PM2.5 pollution, variability, and health impact. *Nat. Commun.* **14**, 8349 (2023).
62. Manubens, N. *et al.* startR: Automatically Retrieve Multidimensional Distributed Data Sets. (2025).
63. R Core Team (2021) R A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. - References - Scientific Research Publishing. https://www.scirp.org/reference/referencespapers?referenceid=3131254.

64. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations - Rue - 2009 - Journal of the Royal Statistical Society: Series B (Statistical Methodology) - Wiley Online Library. https://rss.onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2008.00700.x.
65. Gasparrini, A. Distributed Lag Linear and Non-Linear Models in R: The Package dlnm. *J. Stat. Softw.* **43**, 1–20 (2011).
66. Wickham, H., Francois, R., Henry, L., Muller, K. & Vaughan, D. A Grammar of Data Manipulation. https://dplyr.tidyverse.org/.
67. Pebesma, E. & Bivand, R. *Spatial Data Science: With Applications in R*. (Chapman and Hall/CRC, New York, 2023). doi:10.1201/9780429459016.
68. Oksanen, J. *et al.* vegan: Community Ecology Package. The R Foundation https://doi.org/10.32614/cran.package.vegan (2001).
69. Besag, J., York, J. & Mollié, A. Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Stat. Math.* **43**, 1–20 (1991).
70. Lowe, R. *et al.* Combined effects of hydrometeorological hazards and urbanisation on dengue risk in Brazil: a spatiotemporal modelling study. *Lancet Planet. Health* **5**, e209–e219 (2021).
71. Watanabe, S. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *J. Mach. Learn. Res.* **11**, 3571–3594 (2010).
72. Kramer, M. R^2 STATISTICS FOR MIXED MODELS. *Conf. Appl. Stat. Agric.* https://newprairiepress.org/agstatconference/2005/proceedings/12 (2005).

Supplementary Materials

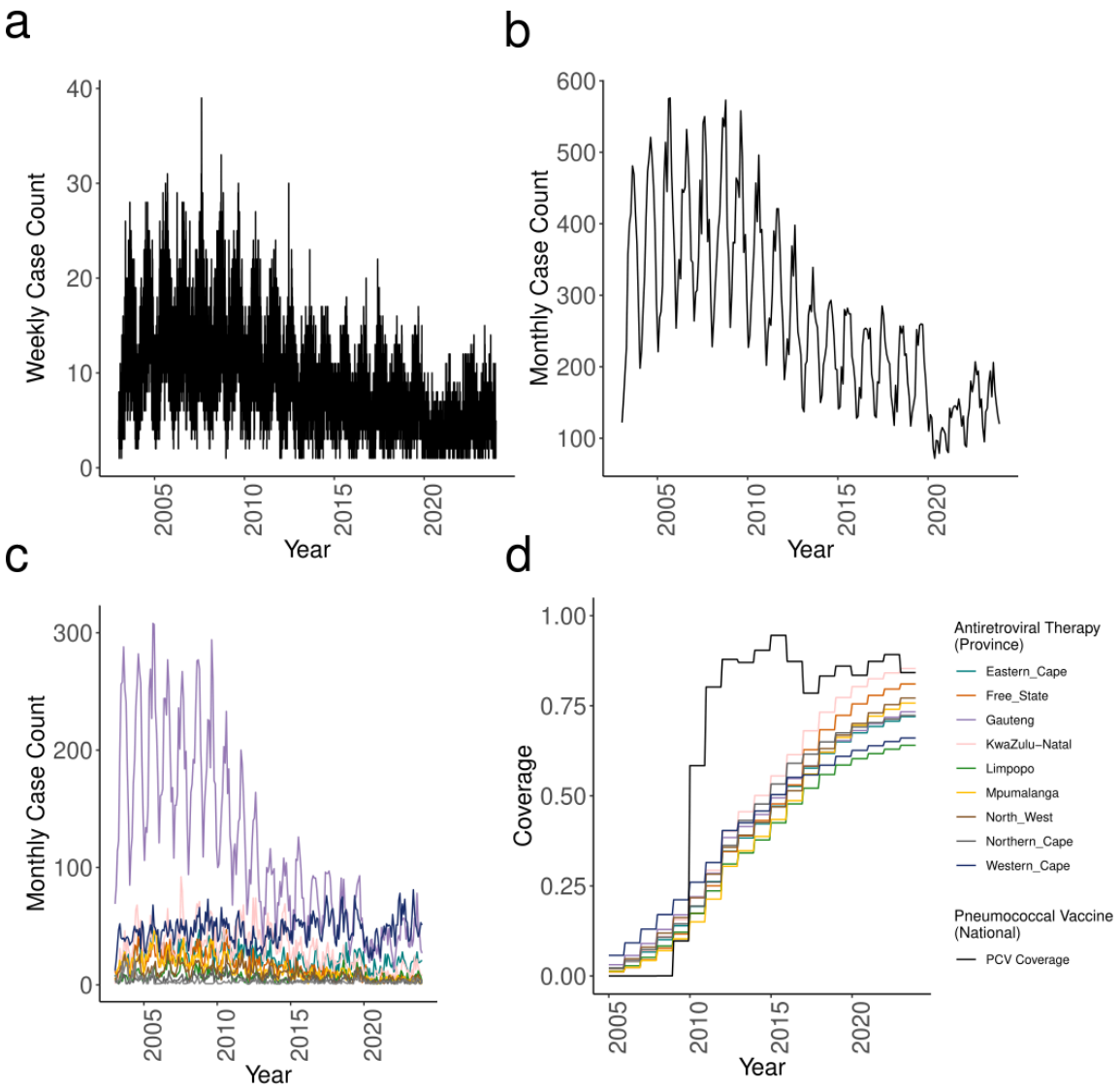

Figure S1. Pneumococcal case counts (a-c) and intervention coverage (d) (a) Weekly (a) and Monthly (b) pneumococcal disease case counts. (b) Disease counts monthly colored by province. (c) pneumococcal conjugate vaccine (PCV) coverage nationally and antiretroviral therapy coverage by province over time.

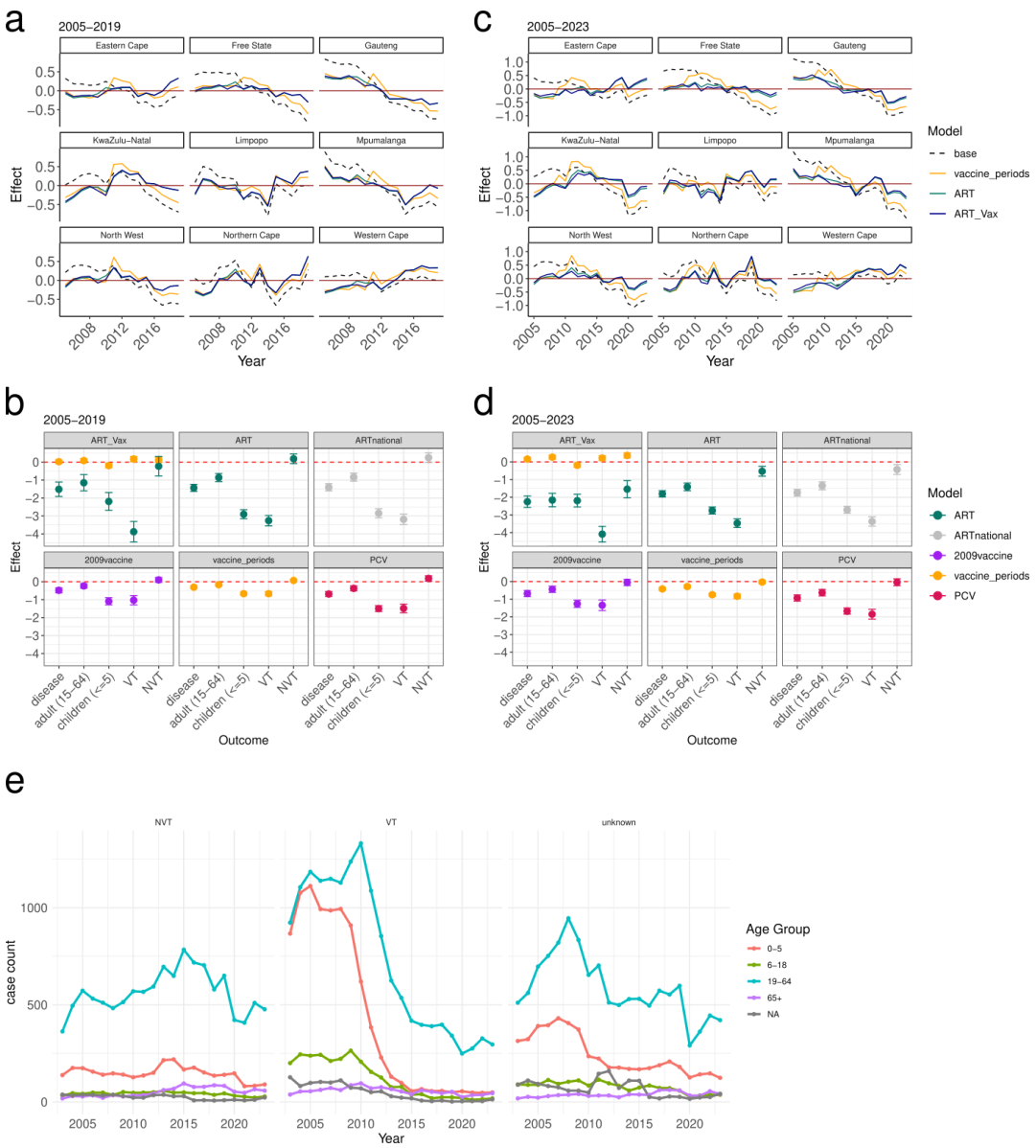

Figure S2. Effects from implementation of ART and vaccine included in the models with years (a-b) 2005-2019 and (c-d) 2005-2023. (a) and (c) show the interannual random effects for the base model with no perturbations (black, dashed), for a model with vaccination periods 2009 and 2011 (yellow), for ART coverage per province (green), and for a model including both vaccination periods and ART coverage per province (blue). (b) and (d) include a model with ART_Vax: both ART coverage per province (green) and vaccination periods (yellow), ART: a model with ART coverage per province (green), ARTnational: ART coverage nationally (grey), 2009vaccine: a model with a categorical variable indicating pre-PCV7 (2005-2008) and post-PCV (2009-2023) (purple), vaccine_periods: a categorical covariate indicating the pre-PCV, PCV7 (2009-2010), and PCV13 (2011-2023) periods (yellow), and a model with PCV coverage nationally (pink). These each were run for outcomes of all IPD, IPD in adults 15-64, children ≤5, VT serotypes, and NVT serotypes independently. (e) shows the case count across years and by age group including 0-5 (pink), 6-18 (green), 18-64 (blue), 65+ (purple), and absent age data (NA) (grey) for disease from NVTs (left), VTs (middle), and unknown serotype (right).

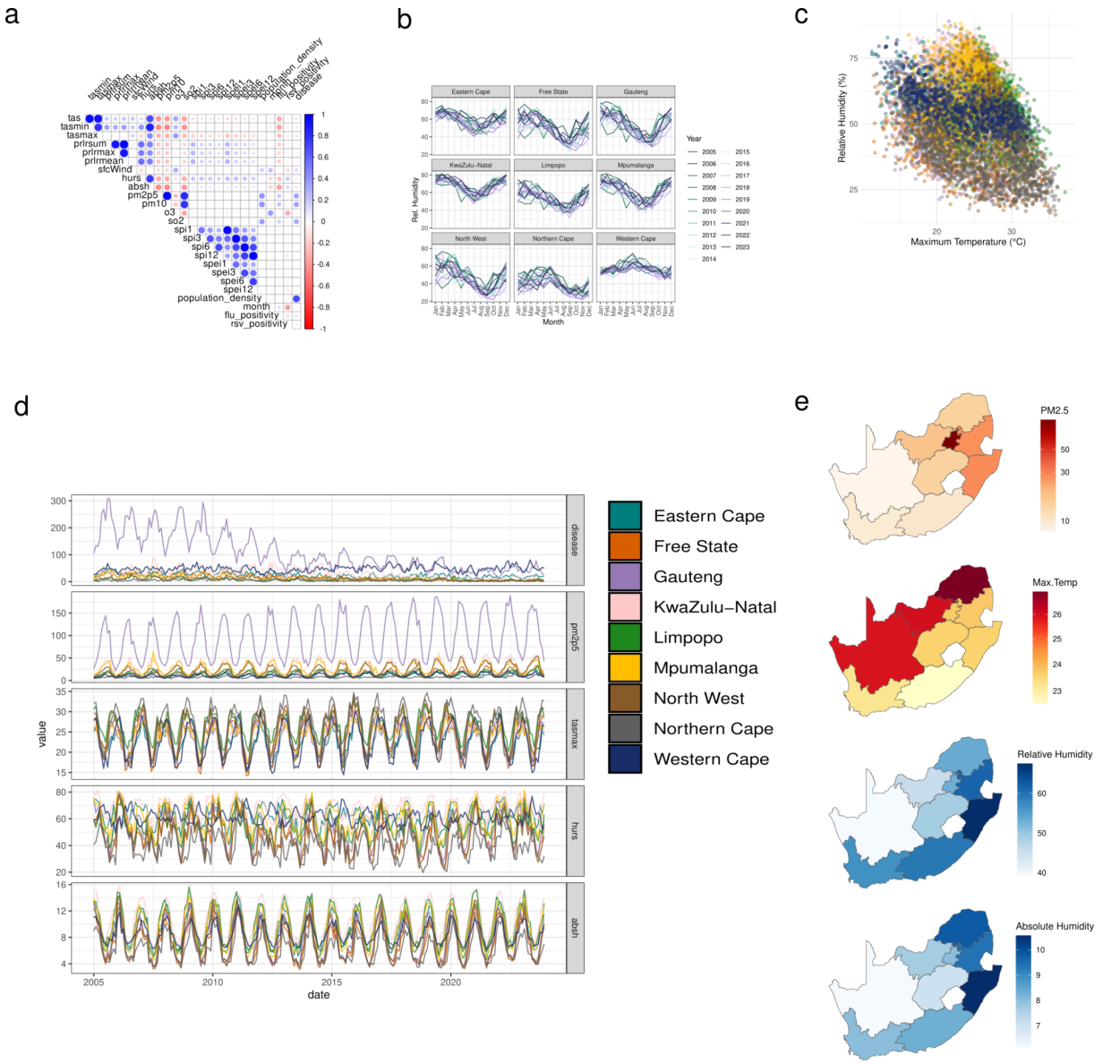

Figure S3. Summaries of included environmental variables. (a) correlation plot with spearman correlation between each of the external variables included. (b) relative humidity across months colored by the years of the study faceted by province. (c) Correlation between maximum temperature and relative humidity colored by province (same legend as d). (d) time series of pneumococcal disease case counts and (e) mean of each environmental variable per the 9 provinces of South Africa (d-e) $PM_{2.5}$ concentration ($\mu g/m^3$), maximum temperature (celsius), relative humidity (%), and absolute humidity (g/m^3 water) across 19 years colored by province.

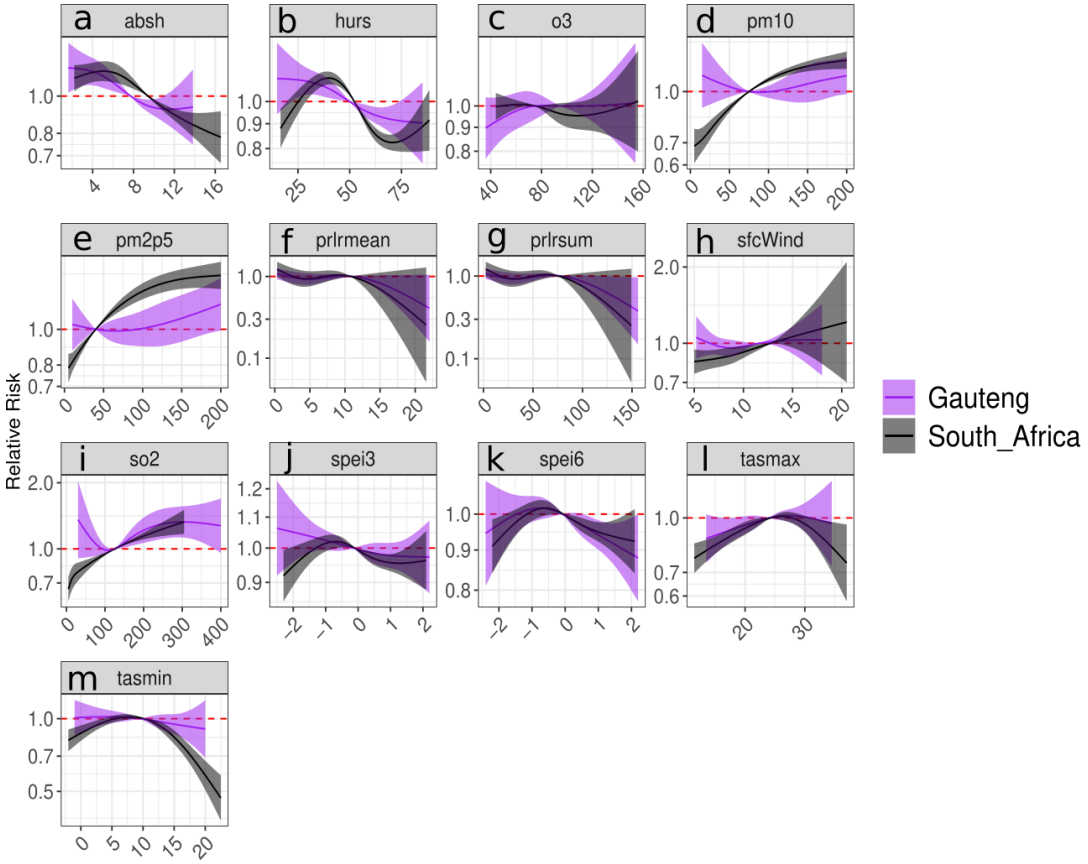

Figure S4. Cumulative exposure-response curves across 8 weeks of lags for the relative risk (y-axis) from each environmental variable (x-axis) in univariable DLNM models for Gauteng alone (purple) and South Africa overall (black). These models are fit to weekly district level case data with the baseline model random and fixed effects (seasonal, spatial, interannual, 3 vaccination period, and population density). (a) absolute humidity (absh) in g/m^3 water (b) percent relative humidity (c) ozone (o3) (d) particulate matter <10µm (pm10) (e) particulate matter <2.5µm (pm2p5) (f) mean precipitation millimeters (g) cumulative precipitation in millimeters (h) mean wind speed in km/hour (i) sulfur dioxide (so2) (j) drought index including evapotranspiration across previous 3 months (k) drought index including evapotranspiration across previous 6 months (l) maximum temperature (celsius) (m) minimum temperature (celsius).

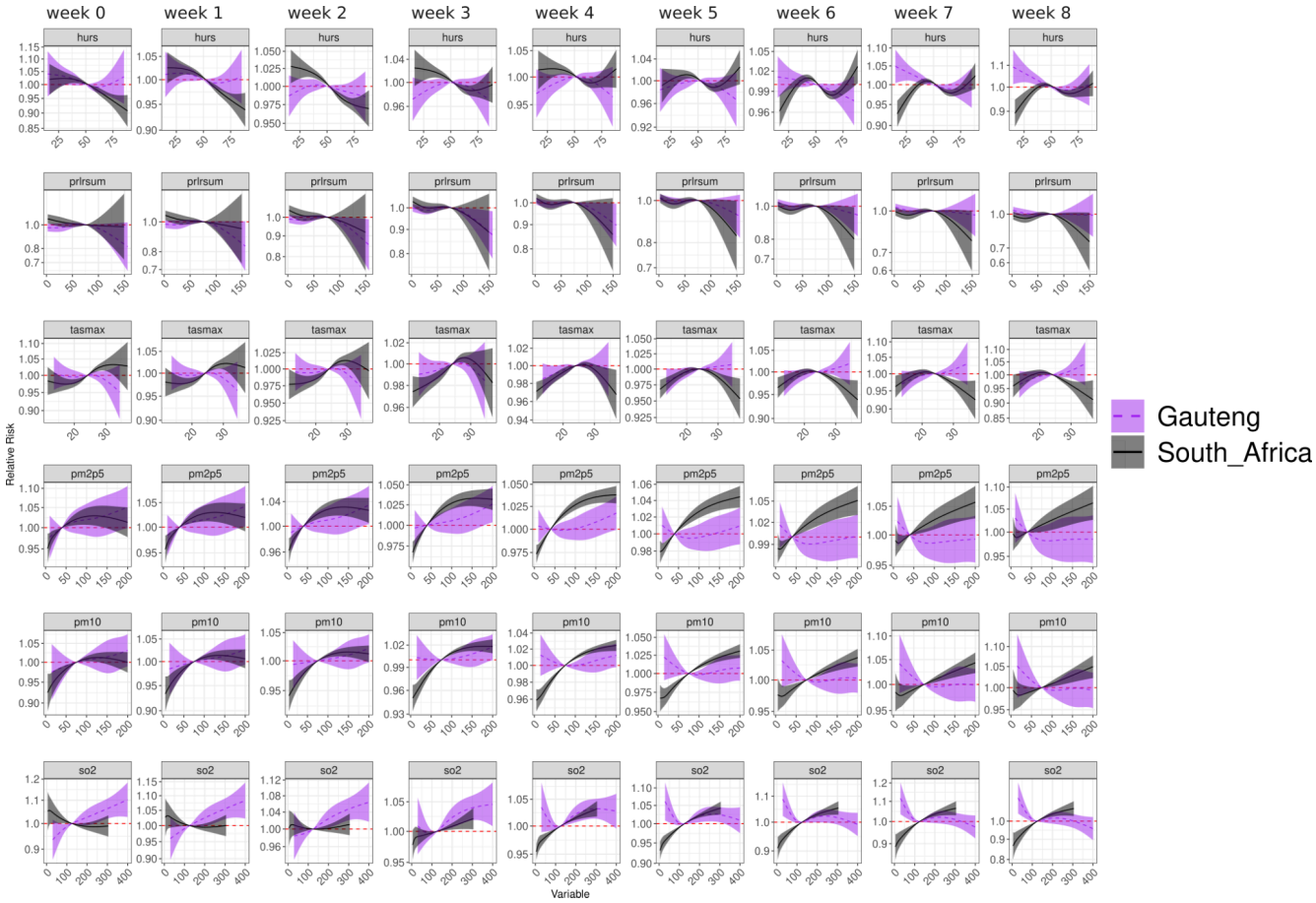

Figure S5. Exposure-response curves across 8 weeks of lags for the relative risk (y-axis) from each environmental variable (x-axis) in univariable DLNM models for Gauteng alone (purple) and South Africa overall (black). These models are fit to weekly district level case data with the baseline model random and fixed effects (seasonal, spatial, interannual, 3 vaccination period, and population density). The metric for each environmental variable is on the x-axis and the relative risk is on the y-axis. From top to bottom these include % relative humidity (hurs), cumulative precipitation (prlrsum) in millimeters, maximum temperature in celsius (tasmax), $PM_{2.5}$ (pm2p5), PM_{10}, and SO_2 concentrations in $\mu g/m^3$.

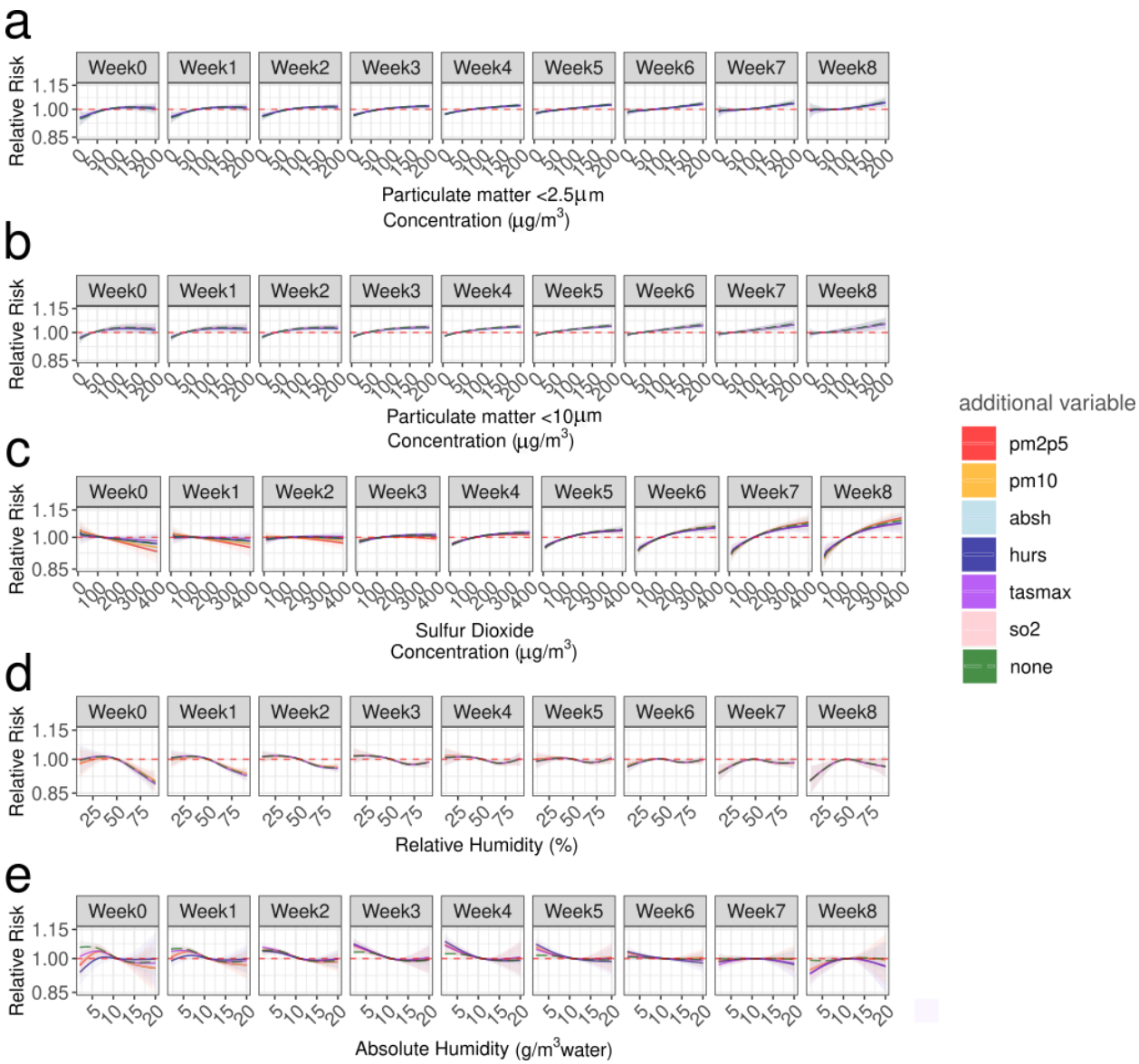

Figure S6. Relative risk across 8-weeks of lags for $PM_{2.5}$, PM_{10} SO_2, relative, and absolute humidity with additional multiplicative variables. In the DLNM the humidity variable cross basis has 3-degrees of freedom (df), while the pollution variables have 2. The additional multiplicative effects are included as fixed effects except the humidity variables which are non-linear with 5 cuts. Each plot includes the exposure-response curve for each lag-week colored by the multiplicative variable for (a) $PM_{2.5}$ (µg/m^3) (models for none, absh, hurs, tasmax, SO_2) (b) PM_{10} (µg/m^3)(models for none, absh, hurs, maximum temperature (tasmax), so2), (c) SO_2 (µg/m^3)(models for none, $PM_{2.5}$, $PM_{10,}$ absh, hurs, tasmax) (d) % relative humidity (hurs) (models for none, $PM_{2.5}$, PM_{10}, SO_2, tasmax) , and (e) absolute humidity (absh) (g/m^3)(models for none, $PM_{2.5}$, PM_{10}, $SO_{2,}$ tasmax).

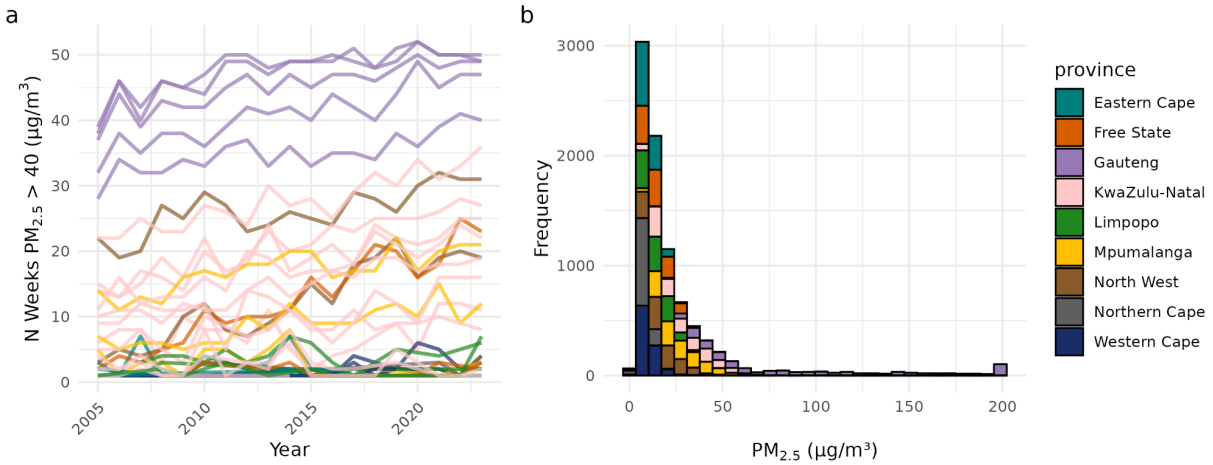

Figure S7. Particulate matter <2.5 µm distributions (a) weeks with >40 µg/m^3 (the NAAQS threshold for $PM_{2.5}$) per year and district colored by province across the time period. The grouping is by district whereby in Western Cape (dark blue) and Northern Cape (grey) have no weeks exceeding the threshold. Eastern Cape (blue) and Limpopo (green) only have districts with <10 weeks (blue), Free State includes Fezile Dabi (orange), Gauteng includes the City of Johannesburg, City of Tschwane, Ekurhuleni, Sedibeng, and West Rand (purple), KwaZulu-Natal includes Amajuba, eThekwini, Harry Gwala, iLembe, Ugu, uMgungundlovu, uMzinyathi, and uThukela (pink), Mpumalanga includes Gert Sibande and Nkangala with >10 weeks exceeding threshold (yellow), North West includes Bojanala Platinum and Dr Kenneth Kaunda with > 10 weeks (brown) (b) histogram of particulate matter distribution across provinces. Both colored by province.

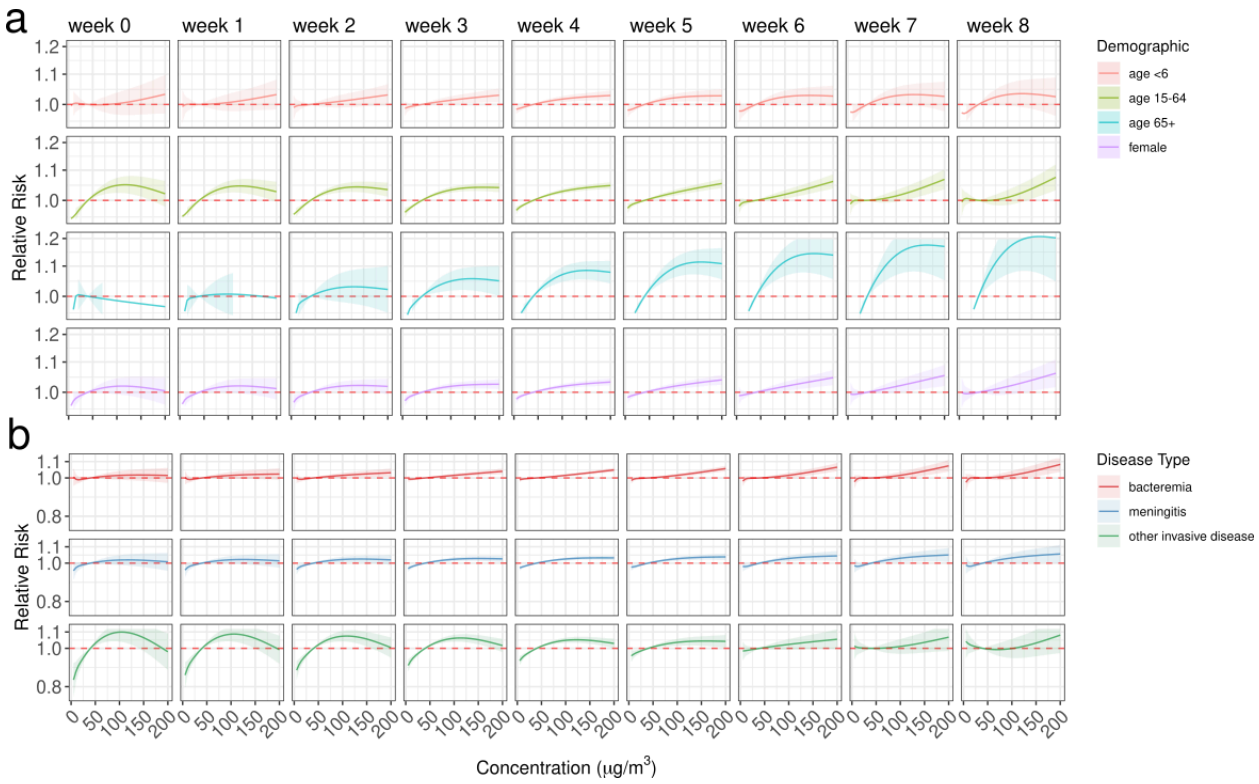

Figure S8. Exposure-response curves across different weekly exposure to $PM_{2.5}$ (concentration μg/m^3) including different stratifications of invasive disease including (a) demographic factors such as age <6 (red), age 15-64 (green), age >64 (blue), and females alone (purple) and for (b) different disease types including bacteremia (red), meningitis (blue), and other invasive disease types (dark green) all across 8 weeks of lags at the weekly district level.

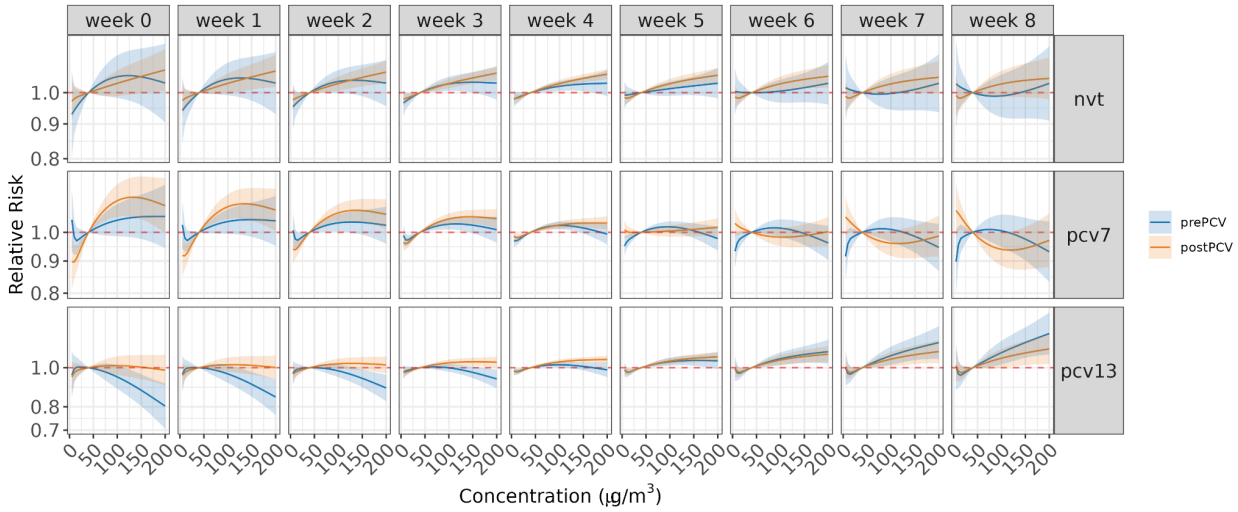

Figure S9. Exposure-response curves across different weekly exposure to $PM_{2.5}$ (concentration µg/m^3) including different stratifications of invasive disease including that which is caused by NVTs (top), PCV7 (middle), or PCV13 (bottom). These are further stratified by the pre-PCV period (2005-2009) (blue) and the post-PCV period (2009-2019) (orange) in a model run at the weekly-province level.

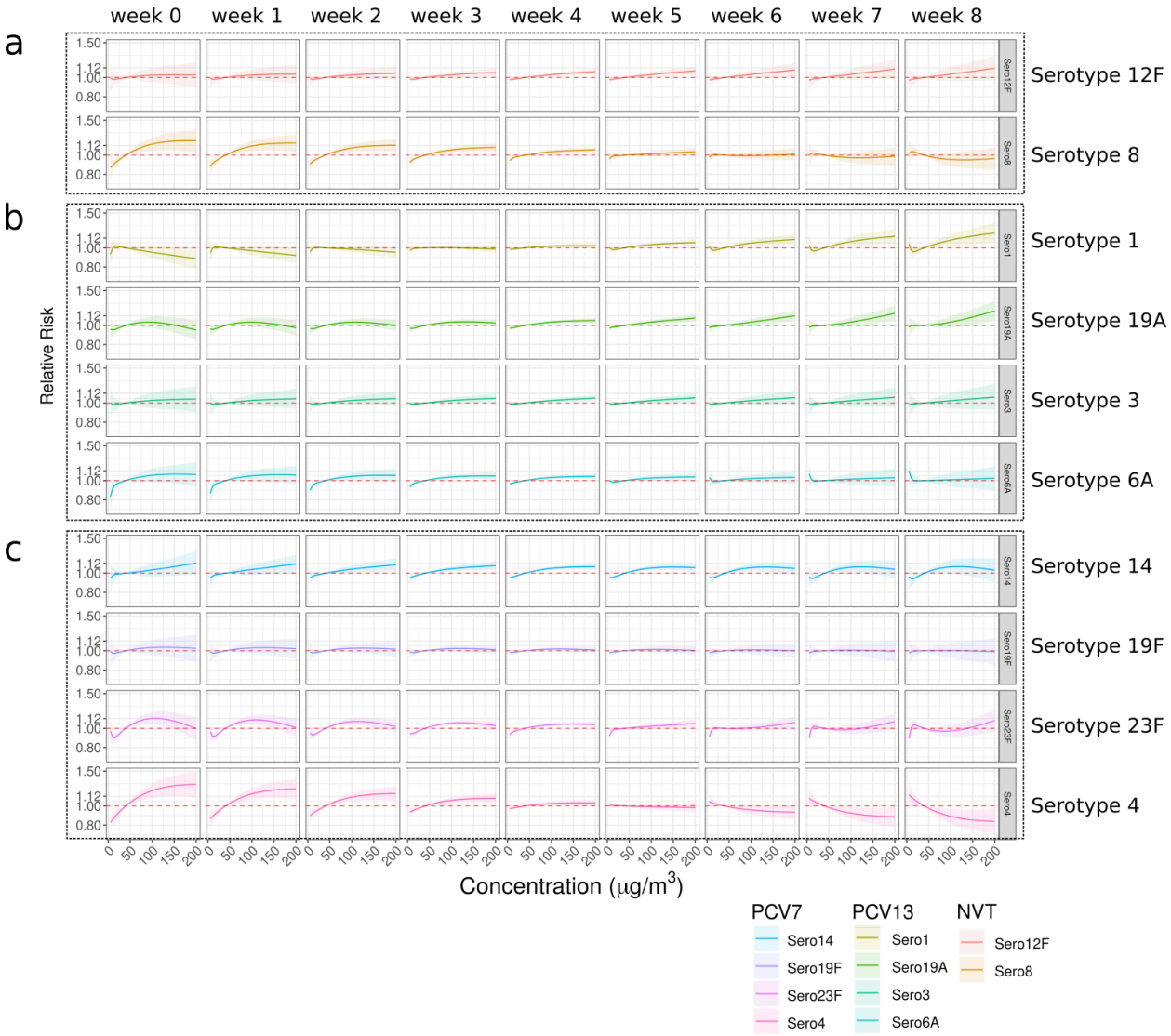

Figure S10. Exposure-response curves across different weekly exposure to $PM_{2.5}$ (concentration μg/m^3) including serotype stratifications of invasive disease including (a) NVTs serotype 12F and serotype 8, (b) PCV13 serotype 1, serotype 19A, serotype 3, serotype 6A, and (c) PCV7 serotype 14, serotype 19F, serotypes 23F, and serotype 4.

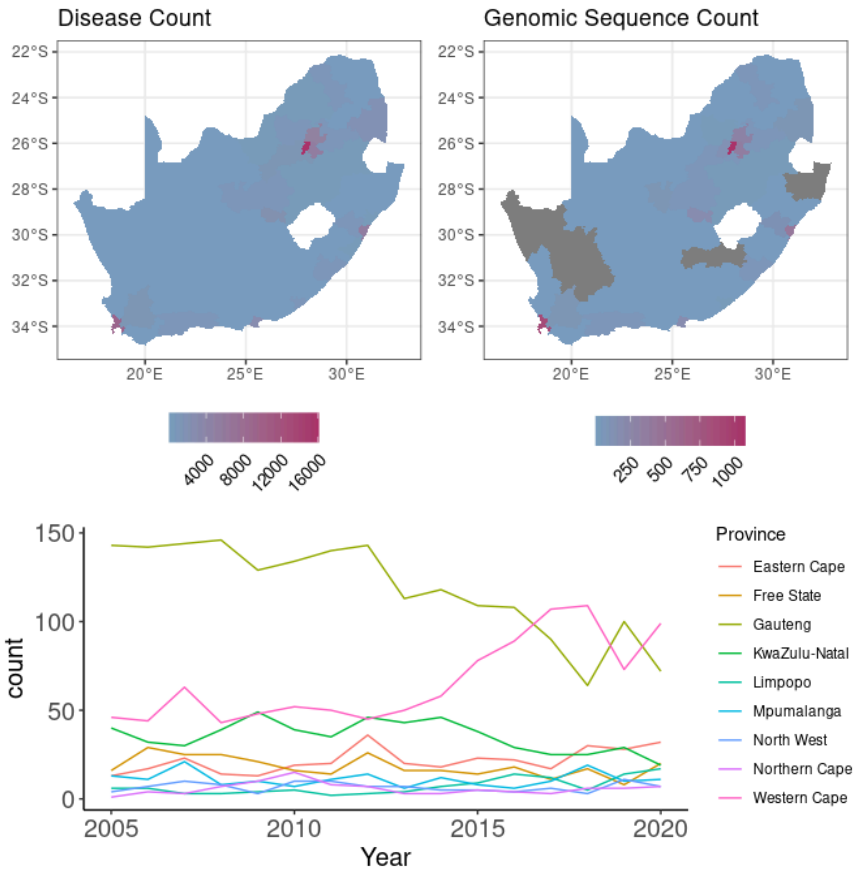

Figure S11. Summary of GPSC sequences as compared to case count and by province (a) Number of disease cases from each district and (b) the number of genomic sequences from each district. There were no sequences from 4 districts including: Namakwa, Joe Gqabi, uMkhanyakude, Zululand. (c) The number of pneumococcal samples sequenced from each province over time.

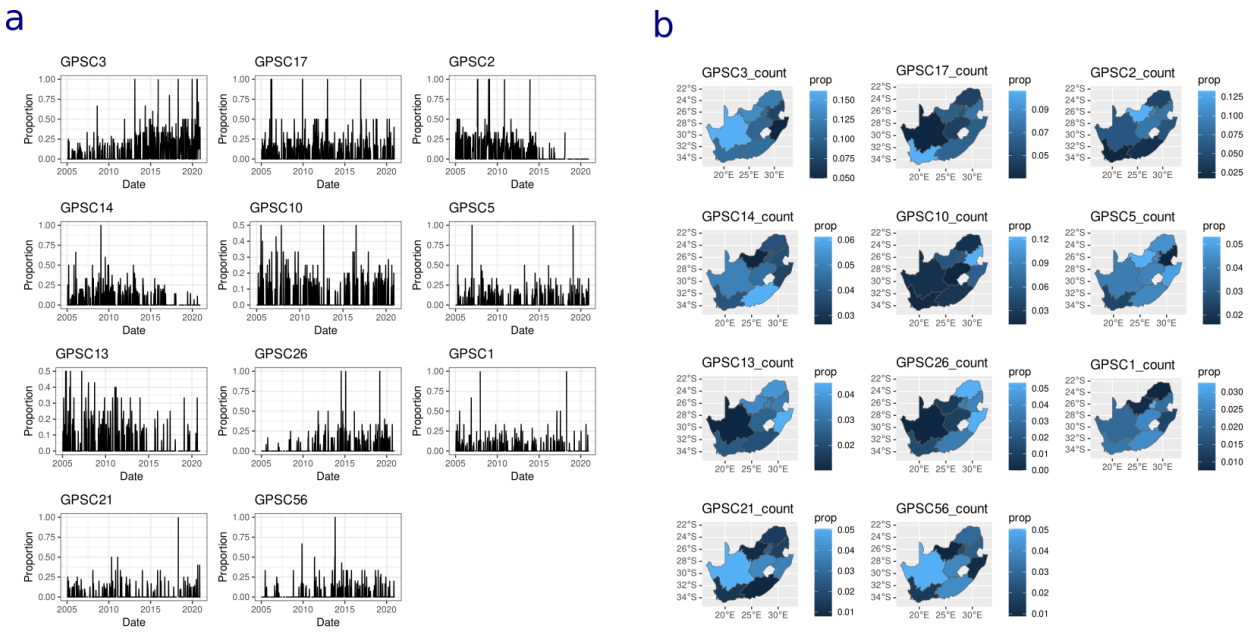

Figure S12. Proportion of each GPSC group annually and spatially per province in South Africa (N=9) across 15 years. The denominator is the number sequenced per week per province.

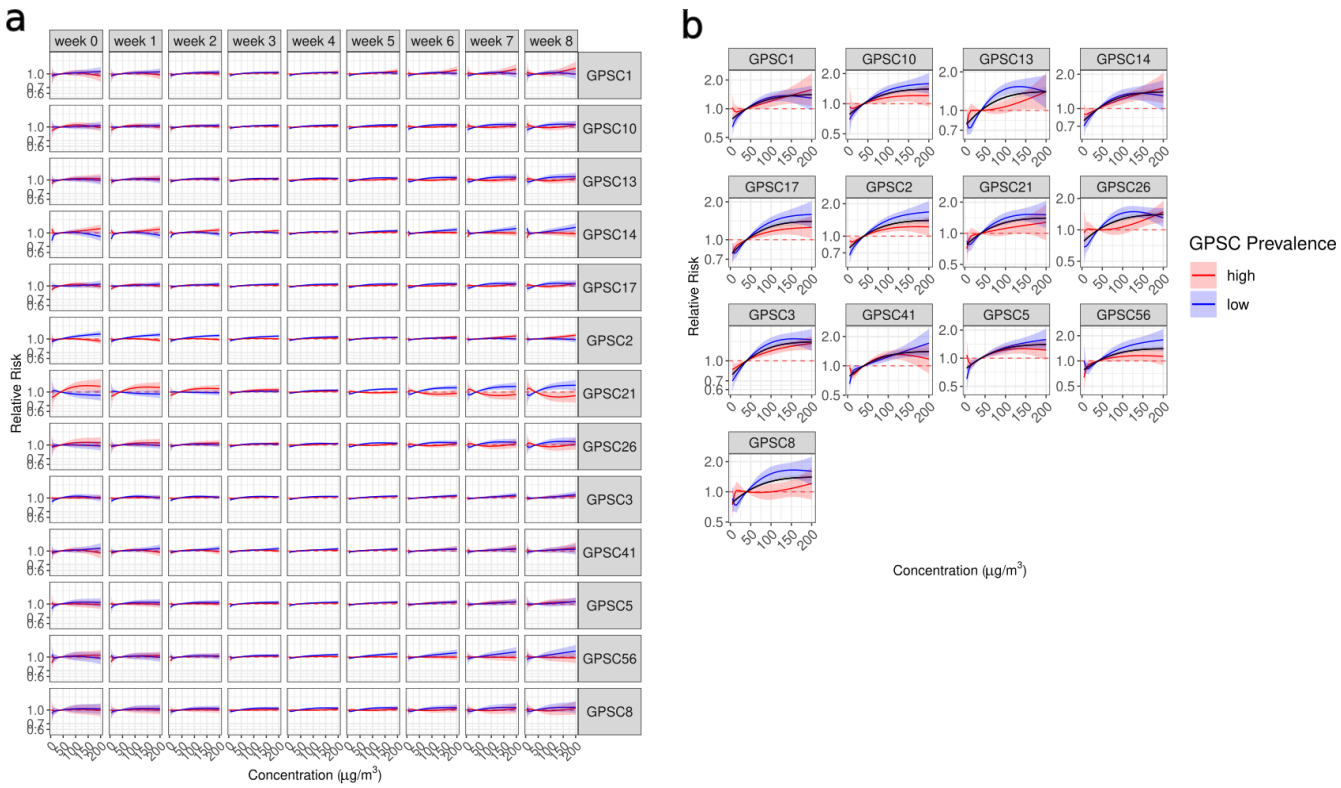

Figure S13. Exposure-response curves for $PM_{2.5}$ and pneumococcal disease interacted with the proportion of each GPSC per week. (a) 8-week lag time exposure-response curves and (b) includes the cumulative effect. The estimates are at high (red) and low (blue) GPSC prevalence. In b the black line indicates the model with no interaction.

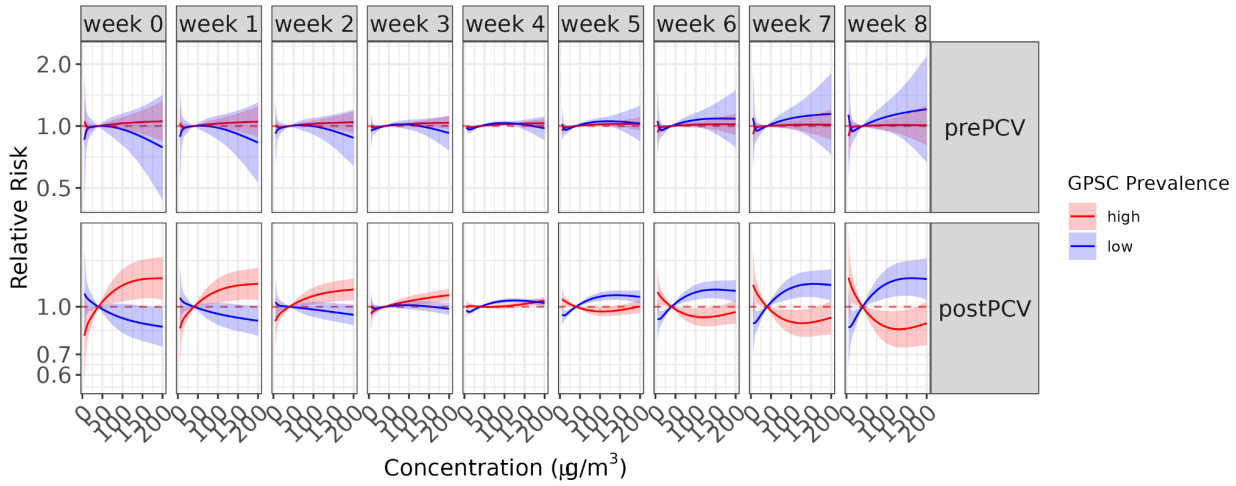

Figure S14. GPSC21-exposure-response curves for the relative risk of IPD given the concentration of $PM_{2.5}$ and interacted with the weekly proportion of GPSC21 in province level weekly models across an 8-week lag for (top) pre-PCV period 2005-2008 and (bottom) post-PCV period 2009-2019.

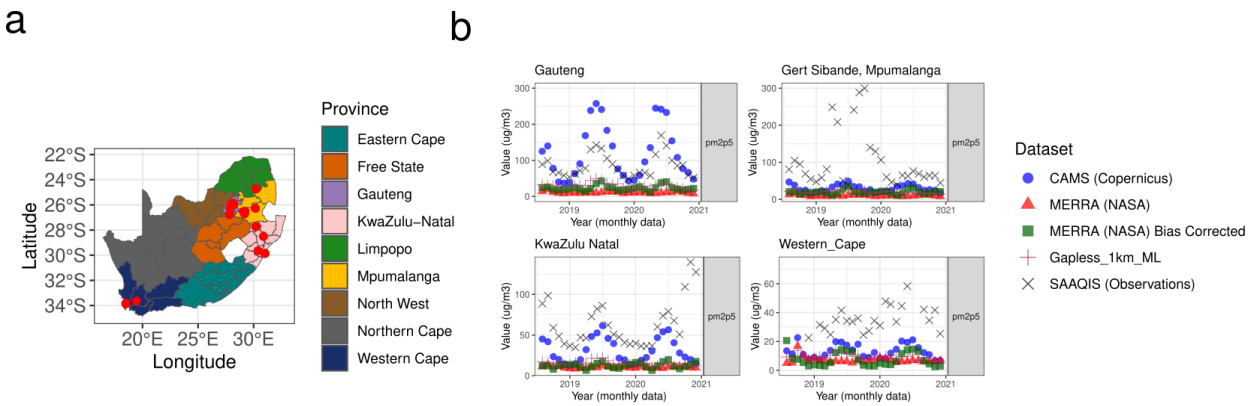

Figure S15. Reanalysis data compared to observations in South Africa. (a) Map of air quality monitoring stations included from SAAQIS. Colored by province with lines distinguishing districts. Locations of observation stations are indicated in red dots. (b) The concentration of weekly $PM_{2.5}$ in $\mu g/m^3$ from Gauteng across reanalysis products from Copernicus (CAMS), MERRA-2 from NASA, MERRA-2 with a bias adjustment and a machine learning gapless product at 1km grid squares.

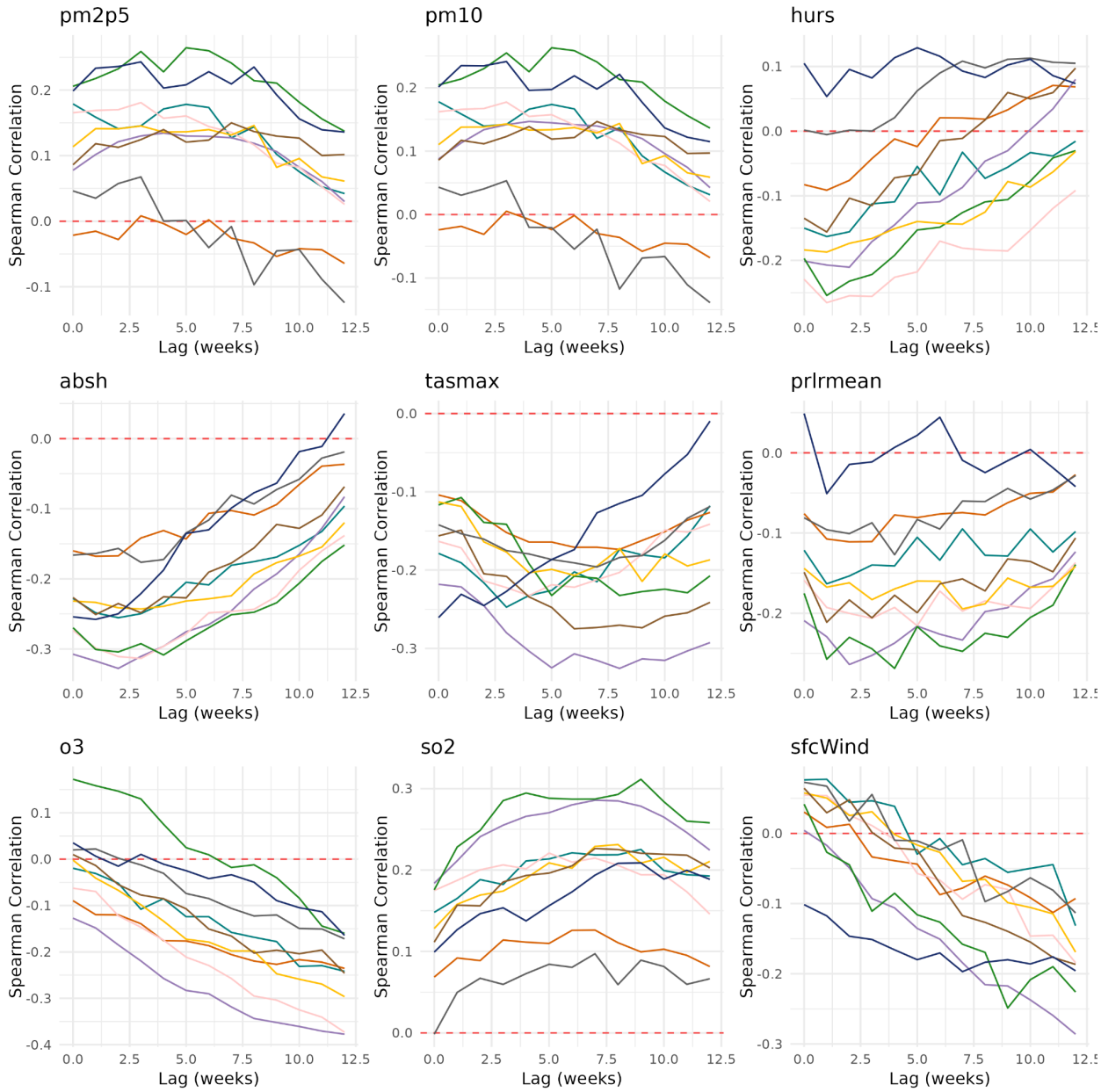

Figure S16. Spearman correlation between each lagged variable across lag weeks with disease case counts for all environmental variables colored by province.

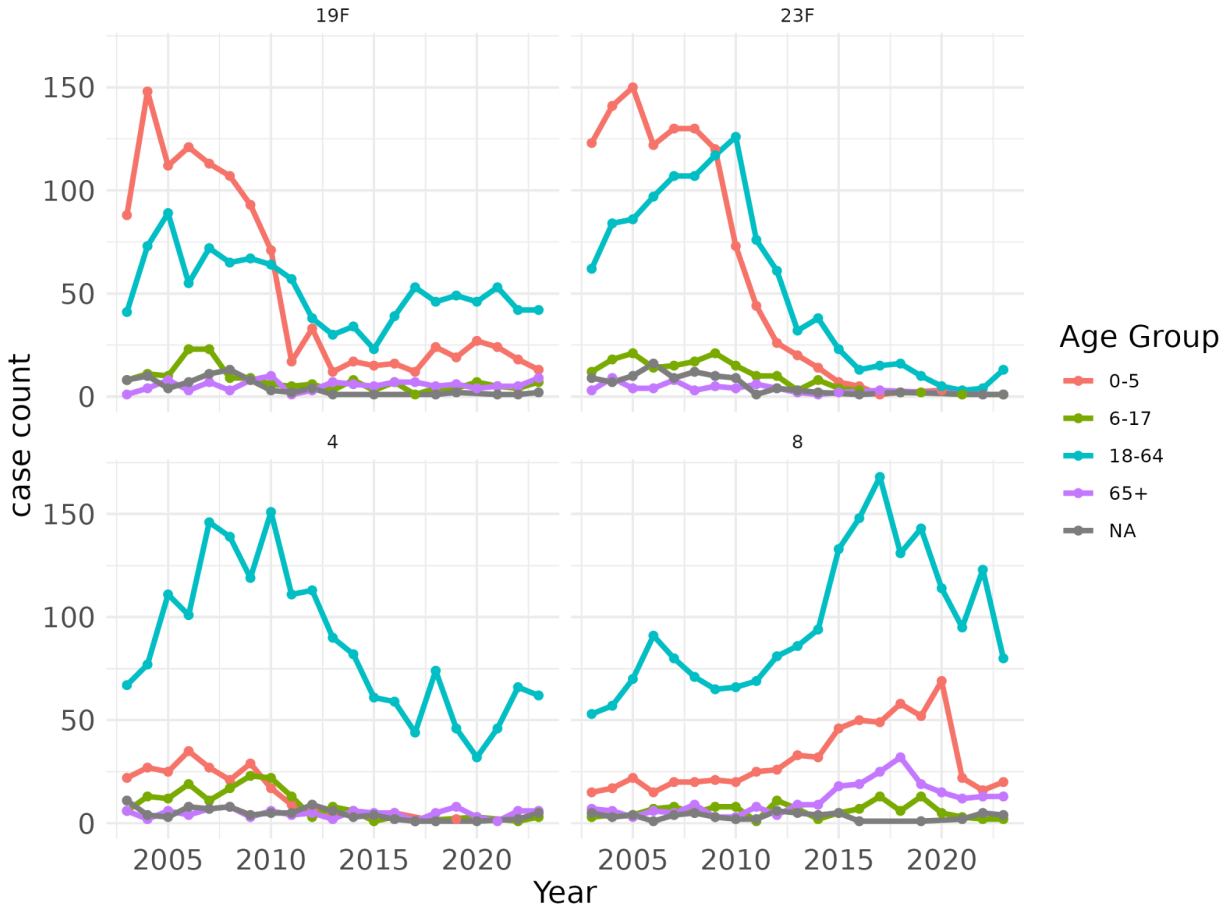

Figure S17. Proportion of IPD cases per age group and year by serotype. Number of IPD cases (y-axis) per year (x-axis) stratified by 0-5 (pink), 6-17 (green), 18-64 blue), and 65 and over (purple). Missing age data (grey). Stratified by PCV13 serotype 19F (top left), PCV13 serotype 23F (top right), PCV13 serotype 4 (bottom left), and NVT serotype 8 (bottom right).