Intellectual Merit: Despite incredible progress in medicine over the past century, the burden of chronic and non-communicable diseases is increasing in many parts of the world [1]. Because genetics do not evolve as rapidly as these changes in health burden, it is hypothesized that environmental and societal changes are driving this increase [1]. This hypothesis, combined with significant leaps in data collection and computational analysis, is generating scientific interest in characterizing the exposome. Theoretically, the **exposome** is the sum total of factors to which each person is exposed over their lifetime, including the environment, lifestyle choices, social stresses and the microbiome [1]. The field of epidemiology offers extensive evidence that certain combinations of these factors are significantly associated with various health risks, and with one another. However, statistically teasing apart the health effects of multiple exposures (which is critical for informing both health care and public policy making) is extremely challenging.

One major problem in quantifying the health impacts of one or more exposures is **exposure measurement error (EME)**, which is the difference between estimated exposure and true exposure. This difference may be an artifact of imprecise data collection (i.e. by sensors or surveys), generalization uncertainty (i.e. from predictive modelling, averaging over an area, etc.), or both. EME often manifests in elevated uncertainty, which often attenuates statistical effect estimates but can also lead to unpredictable bias in certain circumstances [2]. Also, in multi-exposure models, correlation in the EME of different exposure variables may statistically assign some of the observed health effect from a harmful exposure to a benign exposure [2]. However, it has been shown through statistical simulation that developing a more accurate exposure estimation model might not always improve health effect estimates, depending on the exposure data sample size and relative complexity of the exposure model [3].

Another major problem is that different exposures are measured at **different spatial and temporal scales**, each of which may or may not be the scale of health-relevant variability [4]. For instance, in the US, socioeconomic variables are often updated every three to five years at the Census tract level, whereas air pollution measurements are most commonly taken daily at individual monitor locations across a city. Nothing in nature dictates that health-relevant variability in social or environmental exposures takes place at these temporal and spatial scales. Combining such exposures at these different scales and with different types and magnitudes of EME may lead to spurious epidemiological findings. However, to my knowledge there are no studies examining the spatial scale of measurement and impacts of EME on health effect estimates for both environmental exposures (specifically air pollution) and social exposures [4].

This represents a major gap in understanding, relevant to a diverse array of academic and societal applications. To fill this gap, **I propose a statistical simulation project to quantify the impacts of differing measurement scale and EME in both air pollution and social exposures on asthma epidemiological effect estimates.** The choices of air pollution and asthma to represent environmental exposures and health impacts are due both to their prominence in the environmental health literature and to my own research background. In brief, the project will consist of using available data to inform fine-scale simulations of **air pollution, social exposures, and asthma hospitalizations** across a major US city (e.g. Los Angeles), which will be considered "truth". I will then run a set of designed experiments of epidemiological analyses that span the space of fine- and coarse-scale exposures and **quantify the uncertainty and bias** in the asthma effect estimates associated with various "data collection" and "modeling" factors.

To inform fine-scale simulation of environmental and social variables, I will use publicly available data (2010-2020) from the US EPA, NCEI, USGS, Census, and American Community Survey (ACS). Environmental variables will include air pollution measurements (EPA AQS), meteorological variables such as temperature and humidity (NCEI land-based sensors) and landcover data (USGS NLCD). Social variables will include income, education, unemployment rate, age, gender, ethnicity, and violent crime rate from the US Census and ACS. In addition, all the graduate programs to which I am applying have access to large and diverse health data sets and capacity for running intensive computation, as well as biostatistics faculty with expertise in environmental health and spatial statistics. After acquiring environmental, health and social data, I will simulate "true" exposures at fine spatial and temporal resolution using techniques from spatial statistics. In general, spatial data simulation techniques can account for both spatial patterning and random variation. This flexibility allows for simulation of many "truth" datasets using a variety of assumptions about spatio-temporal autocorrelation and correlation between air pollution, social exposures, and asthma hospitalizations, informed by the air pollution and asthma epidemiology literature. I expect that completing the simulations will take my *first year*.

To imitate the different ways that environmental and social exposures and health data are typically measured, I will **down-sample** in space and time by taking environmental data from randomly-selected locations (to represent air quality monitor placement) and aggregating social measures (Census data are often reported as percentages) and summing asthma hospitalizations (population health data are often reported as counts or rates) at the neighborhood, ZIP code, Census tract and county levels. At each level of exposure sampling, I will also add varying amounts of measurement error (both uniform and biased) to simulate imprecise data collection by sensors and surveys. Simulating EME in air quality measurements is especially important in evaluating the scientific research potential of **low-cost sensor networks** (such as the new network that I am helping the Denver Department of Public Health & Environment to calibrate), which may have a higher density of less-accurate sensors. Similarly, simulating EME in social exposures could be considered a representation of using **nontraditional data sources** such as social media posts (which are increasingly being used in scientific research applications) to represent social exposures in epidemiological studies.

In my *second year*, I will use standard epidemiological models to calculate the asthma effect associated with air pollution and social exposures given each combination of original and down-sampled data. This will allow for calculation of the differences in magnitude and direction of the asthma effect estimates and those of the spatiotemporal correlations assumed in the simulation step. In my *third year*, I will employ a variety of statistical techniques (such as post-hoc cluster analysis) to assess the impacts of each modeling and sampling decision on these differences. **The result will be quantification of uncertainty and bias in asthma effect estimates resulting from differences in measurement scale and EME in air pollution and social exposures.**

Broader Impacts: Simulation allows for investigation of many more possible combinations of exposures and health outcomes than observational studies. Especially when considering multiple exposures across measurement scales, simulation is the most feasible way to truly understand the impacts of different levels of uncertainty and bias in data collection, exposure estimation, and associated epidemiological findings. The results from this project will help **inform scientific standards** for data collection and statistical analysis not only in environmental health and the emergent field of exposomics, but also more generally in geography and applied statistics. Finally, quantifying tradeoffs between environmental and social data quantity and quality in health impacts analysis could help **inform public and private investment** in sensor networks, programs for collecting social data, and the creation of more comprehensive health databases.

References: **[1]** Renz *et al., J. Allergy Clin. Immunol.* (2017) **[2]** Zeger *et al., Environ. Health Perspect.* (2000) **[3]** Szpiro *et al., Epidemiol. Camb. Mass*. (2011) **[4]** Humphrey *et al., Int J Env. Res Public Health*. (2019)