Fall 12-16-2013

Modeling Fecal Bacteria in Oregon Coastal Streams Using Spatially Explicit Watershed Characteristics

Paul Bryce Pettus
Portland State University

Modeling Fecal Bacteria in Oregon Coastal Streams Using Spatially Explicit Watershed Characteristics

by

Paul Bryce Pettus

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Science
in
Environmental Science and Management

Thesis Committee:
Yangdong Pan, Chair
Eugene Foster
Melissa Lucash

Portland State University
2013

ABSTRACT

Pathogens, such as *Escherichia coli* and fecal coliforms, are causing the majority of water quality impairments in U.S., making up ~87% of this grouping's violations. Predicting and characterizing source, transport processes, and microbial survival rates is extremely challenging, due to the dynamic nature of each of these components. This research built upon current analytical methods that are used as exploratory tools to predict pathogen indicator counts across regional scales. Using a series of non-parametric methodologies, with spatially explicit predictors, 6657 samples from non-estuarine lotic streams were analyzed to make generalized predictions of regional water quality. 532 frequently sampled sites in the Oregon Coast Range Ecoregion, were parsed down to 93 pathogen sampling sites in effect to control for spatial and temporal biases. This generalized model was able to provide credible results in assessing regional water quality, using spatial techniques, and applying them to infrequently or unmonitored catchments. This model's 56.5 % explanation of variation, was comparable to other researches regional assessments. This research confirmed linkages to

land uses related to anthropogenic activities such as animal operations and agriculture, and general riparian conditions.

ACKNOWLEDGEMENTS

Without the patience of my wife, Tatiana, none of this would have ever been possible. Thankfully, I also had the wonderful support from my parental units, John and Peggy, and my amazing siblings, Ryan, Trevor, and Susan.

Gracious recognition needs to be given to my committee. Gene Foster at the Oregon Department of Environment Quality, for letting me shadow his team through numerous special projects. Melisa Lucash, for her special appreciation and application of the teaching and learning process. Dr. Pan and his lab group, for their dedication to training the next generation of scientists.

Special of thanks to Zoe Rodriguez del Rey, for her pressure to apply to this graduate program, and then guiding me through graduate school.

I also need to thank my friends and mentors at the Oregon Department of Environment Quality and the Environmental Protection Agency. Kevin Brannan, for mentoring me through the rigors of water quality modeling and applied science. Ryan Michie, for his frequent help with GIS processes and navigating ODEQ's data bases. Jonathan Halama, thanks for

being a sounding board and for the many long hours of talking about spatial modeling.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

For decades, fecal indicator bacteria have been used to assess water quality for pathogen contamination and violations of state and federal water quality criteria to protect designated uses (ODEQ, 2010). *Escherichia coli* (*E. coli*) and fecal coliform are often used as indicator bacteria, and compromise the largest group of pollutants that are threatening or causing water quality impairments in the U.S. (USEPA, 2012a). All water bodies within the U.S. that have been tested are to be reported by the states to the Environmental Protection Agency (EPA) for all water quality criteria excursions as required by Sections 305(b) and 303(d) of the Clean Water Act. However, only 27% of river and stream miles have been reported on by states (USEPA, 2012a). Of this subset of tested stream and river waters, 54% of them are either listed as threatened or impaired for one or more water quality criteria. Pathogens, such as *E. coli* and fecal coliforms, make up ~87% of these impairments, making them the largest impairment group (USEPA, 2012a). Public heath can be protected through efficient detection and prediction of indicator bacteria, but unfortunately even the most modern water quality models and methods are limited by the characterization of the watershed, and the particular

processes within a specific basin (Ferguson et al., 2003; Jamieson et al., 2004; Benham et al., 2006; Pachepsky et al., 2006; Oliver et al., 2009). With the majority of water bodies in the U.S. being either in violation of current standards or completely untested, generic regional cross-section models that predict fecal contamination would greatly aid natural resource managers in protecting public health (Smith, 1997; Pachepsky et al., 2006; Kay et al., 2010; Crowther et al., 2011).

Predicting and characterizing source, transport processes, and microbial survival rates is extremely challenging, due to the dynamic nature of each of these components (Jamieson et al., 2004). Point sources such as wastewater treatment facilities are highly regulated for bacteria count effluence, but regulating non-point sources is difficult because livestock and wildlife manures vary greatly depending on animal type and application rate (Jamison et al., 2002). Concrete knowledge on the survivability and transport of indicator pathogens is also confounded by a number of environmental factors, such as soil moisture content and the pollutant's ability to move overland to streams (Desmarais et al., 2002; Mossaddeghi et al., 2008). Efforts in waste water treatment and source control have

greatly reduced fecal contamination in both urban and rural areas, however, many streams remain in violation of water quality standards. Treatment, elimination, and control of microbial contamination from point sources are much easier to accomplish than from disperse non-point sources. Regardless, water bodies that have been tested for indicator bacteria and are in violation of State or federal criteria, leads to a waterbody being listed on the EPA's 303(d) list. After which, a Total Maximum Daily Load (TMDL) is developed for the "impaired" waterbody. Some of the best solutions that meet the needs of TMDLs are developed from complex process based models which incorporate source characterization and future water quality protection (Pachepsky et al., 2006). State of the art mechanistic models are limited by their ability to accurately describe life cycles and loading of bacteria, hydrologic processes, climate conditions, and other physical factors that influence fecal contamination in streams (Sadeghi & Arnold, 2002: Benham et al., 2006 ; Kim et al., 2007). For instance, two widely used mechanistic models, Soil and Water Assessment Tool (SWAT) and Hydrological Simulation Program—Fortran (HSPF), use profoundly different methodologies to simulate processes like manure release and hydrology (Chin et al., 2009). Even though process-based

models are the best tools water-quality managers have, empirical and statistical exploration of pathogen relationships to environmental variables can assist in their development and deployment (Crowther et al., 2010; Wilkinson, 2010; Wilkes, 2011)

Simple statistical loading models can't embody complex loading, fate, transport, and timing processes that mechanistic water quality models can (Wilkinson, 2010). They can however advance the knowledge and understanding of environmental factors that drive contaminant loading and fecal indicator violations (Kay et al., 2010; New Zealand Ministry for the Environment, 2010). Kay et al. (2010) used empirical models to determine source appointment between agricultural and sewage source of fecal indicator violations. The New Zealand Ministry for the Environment (NZME) (2010) also used statistical modeling to understand watershed characteristics that influence fecal indicator violations. Many other people and organizations are turning to empirical and other black box modeling tools, used to explore unknowns in the structure of the data and to interpret pathogen sources in relation to stream water-quality (Wilkes et al. 2009; Crowther 2011; Hevesi et al., 2011). These modeling tools use several methods to generalize a watershed's ability to

have pathogen contamination, or to predict specific bacteria counts of unmonitored or infrequently sampled streams. These statistical functions are derived from spatially-generated watershed variables, instream physicochemical factors, geology, geography, hydrology, and other anthropogenic and land use variables that are known to influence pathogen content (Wilkes et al., 2009; NZME, 2010; Crowther, 2011; Hevesi, 2011).

Oregon is not unique in its need for understanding the role environmental and other factors relate to violations of fecal indicator organisms, but it is unique in its regional characterization of those variables. In 1988 the Oregon Department of Environmental Quality (ODEQ) set out to devise a strategy to prioritize the state's water bodies based in part on ecoregions (Clark et al., 1991). These researchers stated that variations in water quality would be better served by recognizing similarities and differences between ecoregions rather than across watershed boundaries. Depending on the size of a delineated watershed, a stream or river may flow through many distinct geology types, vegetation, and other natural phenomena that vary greatly from start to finish. These differences in ecoregions fundamentally affect water quality. Therefore transferring

already developed water quality models between different regional watersheds is not possible. It is difficulties like these that arise when deriving modeling inputs and characterizing the fate and transport of pathogen contaminants such as fecal bacteria within an unspecified watershed. But, generalized regional statistical modeling techniques such as those used by the U.S. Geological Survey (USGS) could be informative and useful in Oregon's quest to solve its water-quality problems (Smith et al., 1997).

The objective of this study is to build upon current analytical methods that are used as exploratory tools to predict pathogen indicator counts across the Coast Range ecoregion of Oregon. This region of Oregon has been the focus of many TMDL's, and ODEQ (2013) is currently implementing several more in the region. Between the year 2000 and 2010, roughly 16,400 water quality samples from 532 stations were analyzed for *E. coli* or fecal coliforms in the coastal range streams of Oregon (ODEQ, 2012). The state of Oregon employs a monitoring plan that is in part probabilistic and site targeted, while volunteer monitoring groups are less random and more targeted. However, both develop high quality data about the conditions of the state's waters. These samples are neither completely random nor spatially comprehensive in

their placement, but a reasonable regional assessment can be made from these data. The gap in knowledge is not in how to apply rigorous TMDL methodology and solutions to water quality issues, but how to address sparse or nonexistent sampling and use cost effective ways to characterize regional water quality based on publicly available data. I hypothesize that water quality violations of in-stream fecal bacteria are a function of land use, natural factors, and other spatial variables in the watersheds. This generic model will include both sources of indicator bacteria and factors that affect concentration, fate and transport within a watershed. These methods can also be used to predict intensity and identify key watershed variables that drive water-quality violations. It is also my goal to help current watershed management to: 1) Identify likely areas of high pathogen bacteria concentrations in watersheds with infrequent to zero monitoring. 2) Develop generalized models that can be used *a priori* to expensive process-based water quality models. 3) Quantify likely impacts of future land-use, land-cover, and population change scenarios.

BACKGROUND

Pathogen Source

Pathogen bacteria, which are found in livestock manures, animal extracts, and humans, are currently causing numerous water quality violations across the world. *E. coli* is a rod shaped, gram negative, enteric bacteria normally found in the intestines of warm blooded organisms. As such, it is used as a general indicator of pathogen contamination in waterbodies (EPA, 2012a). Watershed sources of fecal coliforms can originate from any combination of urban, agricultural, residential, and natural origins (Figure 1). From these sources, pathogens are then transported either directly or indirectly into streams via point source discharge, disperse overland flow, or direct deposition. Conceptually, we might be better served by visualizing these inputs as either direct or indirect in nature, rather than the regulatory definitions of point and non-point sources to stream entry. Direct contaminant deposition into a waterbody is possible through: agricultural livestock, wildlife, pets, human recreational activity, and rural and urban sewerages. Other more easily accounted for direct sources of bacterial contamination are: combined sewer overflows, wastewater treatment plants, and permitted effluence. While residential septic tanks and

straight pipes are more difficult to assess. Indirectly, pathogens from these same generalized sources may be transported overland by hydrological related processes (Figure 1). It is these non-point sources of pollution that makes prediction and characterization of pathogens difficult.

In Paul and Meyer's (2001) frequently cited review on urban streams, they noted the difficulties in characterizing both point and non-point sources of bacterial contamination. Under baseflow conditions, the USGS found that the Platte River near Denver, Colorado, waste water treatment plants (WWTP) contributed 69% to the river's total flow (Dennehy et al., 1998). Other studies showed that storm events have increased instream bacteria counts 10 fold, and that storm drain sewers and stormwater had both human and animal fecal coliforms (Paul & Meyer, 2008). Genetic ribotyping is becoming a more common way of distinguishing sources of pathogen contamination. Wu et al. (2011) found spatial and temporal patterns in both human and wildlife sources of bacteria. Residential areas had higher levels of human bacteria, while open areas were dominated by wildlife sources. Genetic source characterization in urban streams also point to many other source types, such as domesticated animals (Paul & Meyer, 2008).

In agricultural lands, the primary source of fecal contamination is from grazing lands and livestock related production (Jamieson et al., 2004). In some rural areas livestock have unabated access to streams, and frequently manures are directly deposited into streams. Bacteria counts in grazing lands have been shown to have 5 to 10 times higher levels of pathogens than non-grazed lands (Doran & Linn, 1979). Confined feeding operations and other livestock operations are often under strict guidelines that regulate storage and disposal of manures, but are sometimes not enforced (K. Brannan ODEQ, personal communication, September 23, 2011). Applied manure sludges to land can create interesting lag times before bacteria are transported, and are highly variable between application sites and across particular watersheds (Meals et al., 2010). A study in Tillamook Bay, Oregon found that the most probable sources of fecal contamination were from dairy operations and ineffective sewage treatment in this rural coastal watershed (Benhard et al., 2002). In addition, Benham et al. (2006) noted that some older homes in rural areas have straight pipes that connect residential sewage directly to streams.

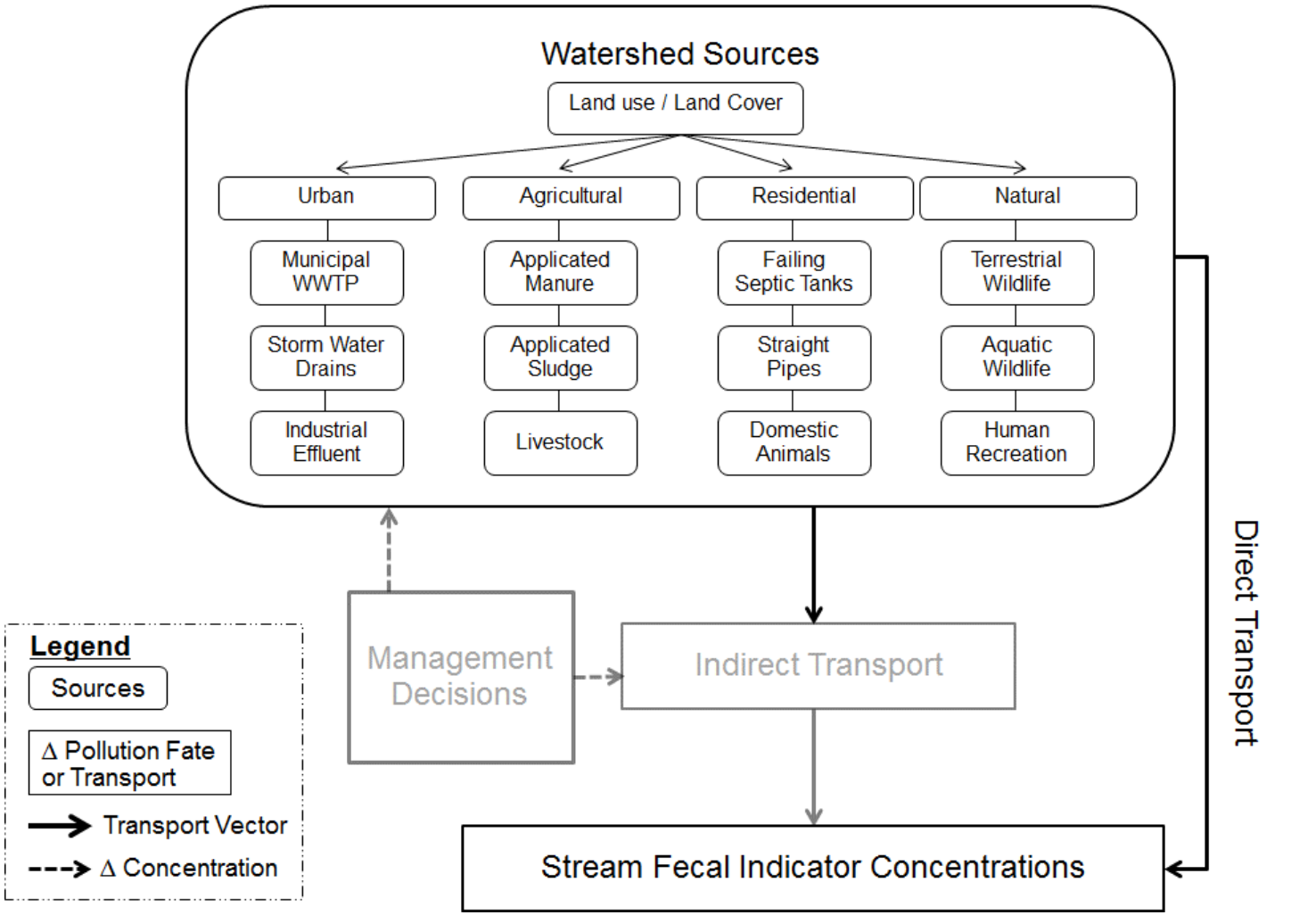


Figure 1. Conceptual model of common watershed sources of in stream *E. coli*.

Pathogen Transport

Indirect fate and transport of pathogens can be categorized in four ways: 1) absorption into soils, 2) migration through soils and into groundwaters 3) overland flow, and 4) bacteria die-off rates (Mossaddeghi et al., 2008). Pronk et al. (2008) warned that water born contaminants can easily be transported through the unsaturated zones of karst aquifers to groundwater networks. While, most researches show that the majority of microorganisms travel by advection in overland flow (Muirhead, 2006). Numerous experiments have been made to study how *E. coli* and other organisms are transported through soil and flow overland. Transport biotracer and artificial biopore experiments, are some of the recent methods to determine the leaching quantity and timing of fecal coliform bacteria in soils (Kuczynska, 2003; Kouznetsova, et al., 2007; Guzman & Fox, 2009; Boyer, et al., 2009). Boyer et al. (2009) used various intact soil samples extracted in the field and returned them to the lab to determine how bacteria move through macroporous soil to the water table (2009). Guzman and Fox (2009) are using artificial biopores to measure pathogen transport interactions between micropore and mesopores. Other researchers are using immunomagnetic

electrochemiluminescence with surface applied biotracers, along with downstream water quality monitoring and down watershed soil sampling to quantify bacteria movement (Abu-Ashour & Lee, 2000; Kuczynska, 2003). Migration of bacteria through soils requires *E. coli* to overcome soil adhesion forces, mechanical pore filtration, and straining through soil mediums (Boyer et al., 2008).

Pathogen Fate

Along with transport studies, other researches show that during the indirect overland transport the fate of bacteria are influenced by a myriad of abiotic conditions and other watershed characteristics (Figure 2) (Table 1). Often it is assumed that bacteria are transported in dissolved solution, as are other non-organic pollutants (Boyer & Kuczynska, 2008; Ponk et al., 2008). These various studies show that most fecal bacteria penetrate only the top 2 cm of the soil, and are almost entirely transported to the stream by surface runoff (Abu-Ashour & Lee, 2000; Kouznetsova, 2007). Overland flow of pathogens to surface waterbodies is affected by both vegetation and the macroporous nature of the regions soil, thus affecting the timing and exposure to environmental factors that influence survival (Boyer & Kuczynska, 2008). Vegetation type and the size of riparian buffers zones will also influence fate and timing to streams. Distance from pathogen source and stream bank slope, in connection with precipitation events will determine timing to stream input and exposure of *E. coli* to abiotic influences (Jamieson et al., 2004). Temperature, extreme dryness, soil moisture, and ultraviolet light have all been shown to affect bacteria transport and life cycles (Boyer & Kuczynska, 2008). Becker

et al. (2010) measured die-off rates of *E. coli* from dairy manure lagoons across a range of temperature treatments. They found that bacteria growth rates increase from 4 °C to 23 °C, and that the *E. coli* die-off sharply as temperatures increase above 23 °C. Differences in geology and soil texture have been shown to influence quantity and timing of stream contamination. Bacteria attached to fine soils, like clays, have higher survival rates than when on coarser sandy soils, and is most probably related to moisture content in the soils (Mubiru, 2000).

Once pathogens have been transported to streams, other abiotic and biotic processes influence their fate (Table 2). Water quality factors such as pH and salinity put osmotic and other stresses on bacteria, reducing their ability to survive (Rhodes & Kator, 1988). Bacteria transported to streams are typically attached to sediments, and resuspension of sediments during high flow events is seen as one of the major issues of increased pathogen counts during these events (Garzio-Hadzick et al., 2010). Garzio-Hadzick et al. (2010) found linkages between water temperature and sediments, showing that bacteria survive better in sediments with cooler waters. Researchers used host-specific bacteria from cows and humans to explore die-off rates in varying sunlight

scenarios, and found that rates were slowed in darkness for both source types (Walters & Field, 2009). Various other factors such as nutrients (NO_3^-, NH_4^+, and PO_4^{3-}) and predation also affect growth and mortality rates once pathogens are in the water (Walters & Field 2009; Williams et al 2012).

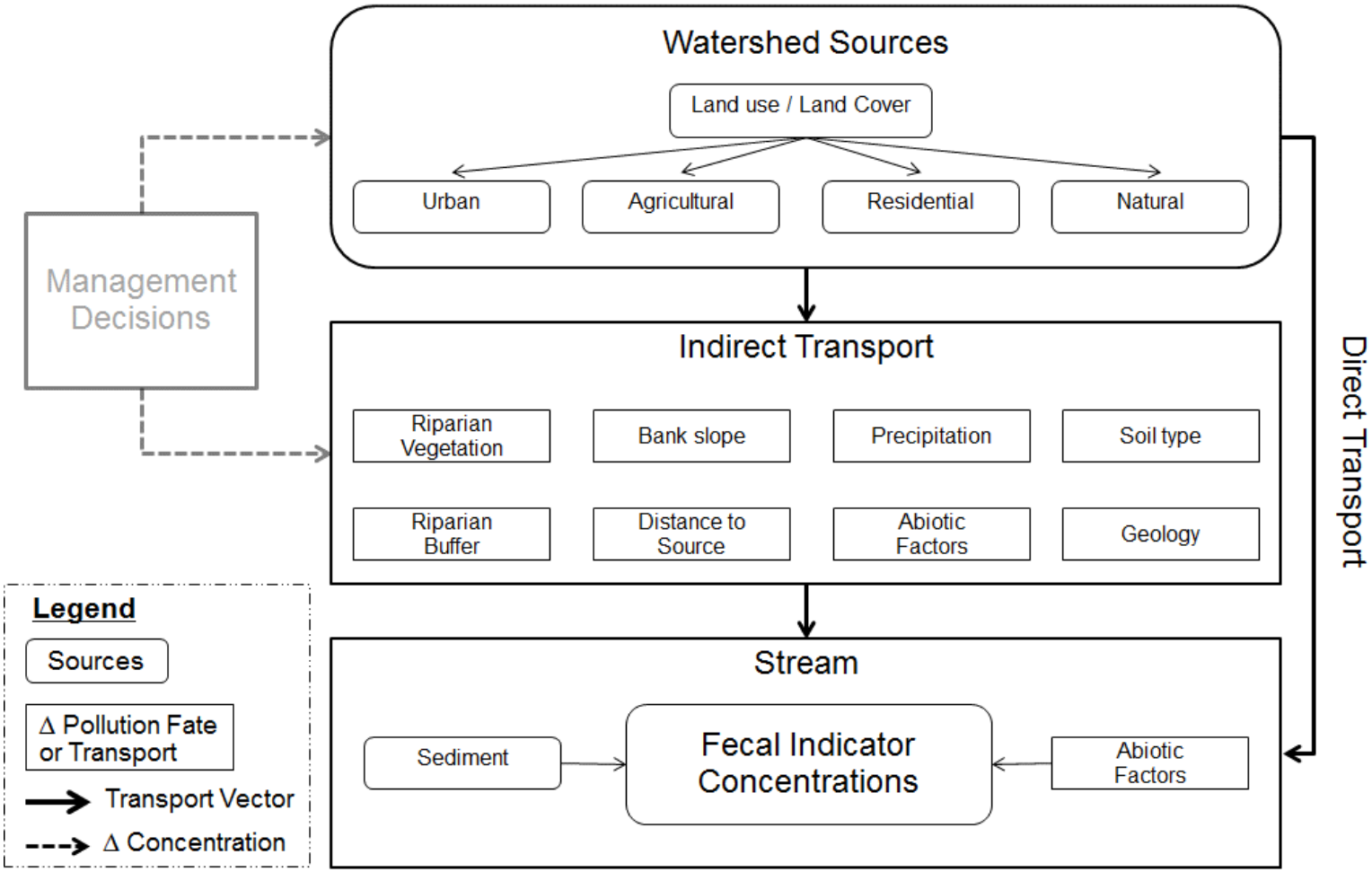


Figure 2. Conceptual model of common factors that influence the fate and transport of *E. coli* in a watershed.

Table 1. Common environmental factors affecting the survival and transport of *Esherichia coli* in a watershed.

Watershed factors	Effect summary	Source
Air temperature	Growth rates of colonies tend to increase in air temperatures from 4°C - 23° F, and fall sharply to temperatures at 40°+.	Becker et al. 2010 Francis & O'Beirne 2001
Humidity/Soil moisture	Wetted organic soils increased survival rates especially after precipitation, and dry soils increased mortality rates.	Jamieson et al. 2004
Soil type	Fine grain soils show lowered colony survival rates than coarser silt soils, but both were influenced by moisture.	Mubiru et al. 2000
Geology	Differences in hydrogeology and aquifer make up influence ground water contamination of *E. coli* counts.	Leber et al 2010
Stream bank slope	High slope conditions increase sediment and nutrient runoff to streams.	Jamieson et al. 2004 Sekely et al. 2002
Sunlight	*E. coli* mortality is highly sensitive to increases in UV radiation.	Gascón et al. 1995
Vegetation / Landuse	Silvopastures had lower bacterial counts in sub surface water than grassland pastures and non-grazed hardwood forests.	Boyer & Neel 2010
Riparian buffer	Vegetative grass buffers can significantly retain *E. coli* from stream entry	Tate et al. 2006

Table 2. Common environmental factors affecting the in stream fate of Esherichia coli in streams.

In stream conditions	Effect summary	Source
Stream temperature	Sediment reservoir *E. coli* have orders of magnitude increased survival rates in cooler 4° C water, than 14° C and 24° C freshwaters.	Garzio-Hadzick et al. 2010
Salinity	Osmotic stress and other abiotic factors increase bacteria die-off in estuarine and intertidal rivers.	Rhodes & Kator 1988
pH	Bacteria have higher mortality in soils and sediments that have lower pH, and survive better in alkaline soils.	Jamieson et al. 2004
Predation	*E. coli* and other allochtonous bacteria are grazed on by protozoa, lytic bacteria, and phages.	Barcina et al. 1997
Sediment	*E. coli* survive longer in stream sediments than in the over laying water, and they become resuspended during storm events.	Garzio-Hadzick et al. 2010
Nutrients	Improved *E. coli* survival is linked to land use and increases in nutrient inputting from runoff of (NO_3^-, NH_4^+, and PO_4^{3-})	Williams et al. 2012
Sunlight	Both human and bovine *E. coli* survive longer in dark microcosms than light microcosms.	Walters & Field 2009

Pathogen Modeling

It is difficult to accurately estimate loading from non-point sources into all waterbodies because of differences in soil types, topography, climate, and land uses. Regardless, water quality managers must develop reasonable models to predict current and future pathogen inputs into streams for specific watersheds. Typically watershed managers use one of many EPA suggested mechanistic models to characterize source inputting, fate process, and potential remediation scenarios (USEPA, 2012b). One of the most widely cited review papers by Jamieson et al. (2004) clearly lays out the difficulties of source characterization, and fate and transport processes that influence enteric bacteria modeling. Besides point source loading, bacterial loading models generally try to model the fate and transport of pathogens via land transport, in stream transport, soil infiltration, storage and movement through the vadose zone, and groundwater hyporheic zone stream entry points (Benham et al., 2006; Kim et al., 2009). Often mechanistic hydrologic models, like SWAT and HSPF, assume that bacteria are transported in dissolved solution, as are other non-organic pollutants (Boyer & Kuczynska, 2008; Ponk et al., 2008). The best of these mechanistic models take into account numerous processes and watershed factors

and must be finely tuned and calibrated to each new project. This setup, calibration, and validation process is extremely time consuming, and therefore expensive. With the need to characterize the probable condition of a state or country's water quality, researchers are developing empirical desktop methodologies to explore water quality in a cost effective manner (Crowther et al., 2001).

In the U.S., the U.S. Geological Survey (USGS) developed a complex spatially referenced regression model which predicts regional water quality (Smith et al., 1997). Smith et al. (1997) developed the SPARROW model to address common problems in assessing regional water quality. Some of the difficulties they stated are: scarce sampling locations due to limited management budgets, focused sampling selection to characterize causes and sources of contamination, and nonuniform basin characteristics between sampling sites. This model linked spatial land use and geographic attributes, hydrology, and source generation to make a regional prediction map of the continental United States using hundreds of monitoring sites and years of hydrological data. SPARROW was able to characterize total phosphorus (R^2 = 0.82) and total nitrogen (R^2 = 0.88) loading to streams and then relate that to infrequently or never sampled streams on a

multi-state regional scale. The authors also conclude that the model gives an understanding to the important factors that affect water quality (Smith et al. 1997). Even though the SPARROW model was not developed for pathogen contaminant transport, the techniques used to statistically analyze how stream nutrients relate to land use and other spatial variables could be informative to other water quality violations such as pathogens.

Other researchers around the world have been using desktop empirical techniques to address nationally mandated water quality policies that are similar to the US Clean Water Act. Researchers in the United Kingdom (UK) are using regression models linking land use type to predict fecal indicator organisms instead of using animal counts, grazing density, and manure application rates (Crowther et al., 2003). In 2003, Crowther et al. used a stepwise procedure to build a multiple regression model linking land use in 20 catchments ranging from 0.7 - 178 km² to *E. coli* counts. The independent variables included land use and basin morphology features such as: % pasture, % woodland, % build up (urban), stream slope, mean altitude, and flow distance. With this model the researchers were able to account for 81.6 % to 82.9 % (R^2) variation in bacteria counts, during low and high

flow periods respectively (Crowther et al., 2003). These UK researchers have been progressing their researches on land use and other geographic data models for source appointment and catchment export coefficients in surface waters, and then exploring land change and best management scenarios (Crowther et al., 2003; Kay et al., 2005; Kay et al., 2008; Kay et al., 2010). More recent researches are now moving towards "Generic Models", which are used to predict or estimate likely pathogen concentration in surface waters across the country (Crowther et al., 2011). These newer regional models are having better results by including population variables such as human and livestock counts along with land use/cover characteristics; this increased the results of previous regional models adjusted R^2 values from 0.54 to 0.62 (Crowther et al., 2011).

In unpublished research, the New Zealand Ministry for the Environment (NZME) used a statistical machine learning method called random forest to make nationwide predictions of *E. coli* stream concentrations (NZME, 2010). Conceptually, the statistical technique these researchers used can be thought of as a type of multiple linear regression, but it is not. Random forest is a type of multivariate non-parametric classification system, which does not rely on the many overlaying assumptions that regression statistics rely on.

Regression statistics assume normal distributions and standardize variability in the data, whereas classification trees or other nonparametric methodologies do not. Ecological and other environmental data, such as bacteria counts and natural factors generally violate these assumptions (Cutler et al., 2007). From 396 spatially diverse sites they used 28 variables that incorporated land cover, climate and flow, and catchment geologic and topography features to model bacteria counts (NZME, 2010). This bootstrapped classification and regression tree model was able to explain ~%70 of the variance of *E. coli* (count/100 ml), with a mean prediction of 256 and a standard deviation of 361 (NZME, 2010). Catchment elevation, % heavy pasture, and rain variability were found to be the most important predictors of bacteria counts in this study. The NZME researchers then used this model to create a prediction map of New Zealand's water-quality in untested or infrequently test surface waters across the country.

These different approaches have a common theme, of taking available water quality data, with likely culprits that affect pathogen loading and fate, to predict surface water quality in rarely or infrequently sampled waters. In my study, I used similar techniques to assess water quality.

METHODS

Study Area

Bacteria sampling station selection was limited to the Oregon portion of the Coast Range Level III ecoregion for reasons related to transferability, regional water quality needs, and data availability (Figure 3). Clark et al. (1991), some of the original contributors to the Oregon ecoregion project, note that by recognizing similarities and differences between ecoregions rather than across watershed boundaries state managers could more effectively assess trends in water quality from point and nonpoint pollution sources. They also state that results from regional assessments could be more reliably extrapolated to a region as a whole when limited by a few number of sampling sites (Clark et al., 1991). As a result of numerous water quality violations, the ODEQ has implemented several bacteria related TMDL's from the northern mid-coast's streams, and is currently developing other TMDL's along the coastal region (ODEQ, 2013a). These pathogen impairments are violating both recreational contact and shellfishing industry use designations (ODEQ, 2013a). Public health managers, the EPA, and ODEQ are especially concerned with source identification and reducing bacteria contamination to an already threatened

shellfishing industry (ODEQ, 2011). Within the region, the use of coastal waters and mountainous streams is common for both angling and recreation. Pathogens exposure to humans from the recreational use designation is more of a concern in the summer dry months from a management and health point of view. Water quality in the Coast Range is the second most frequently sampled ecoregion, after the Willamette Valley (ODEQ, 2012). Due to the health concerns over toxic shellfish and pathogen exposure to recreational users, frequent sampling in the region, and continued and extensive focus from water quality managers and stakeholders, the Coast Range made for a prime case study.

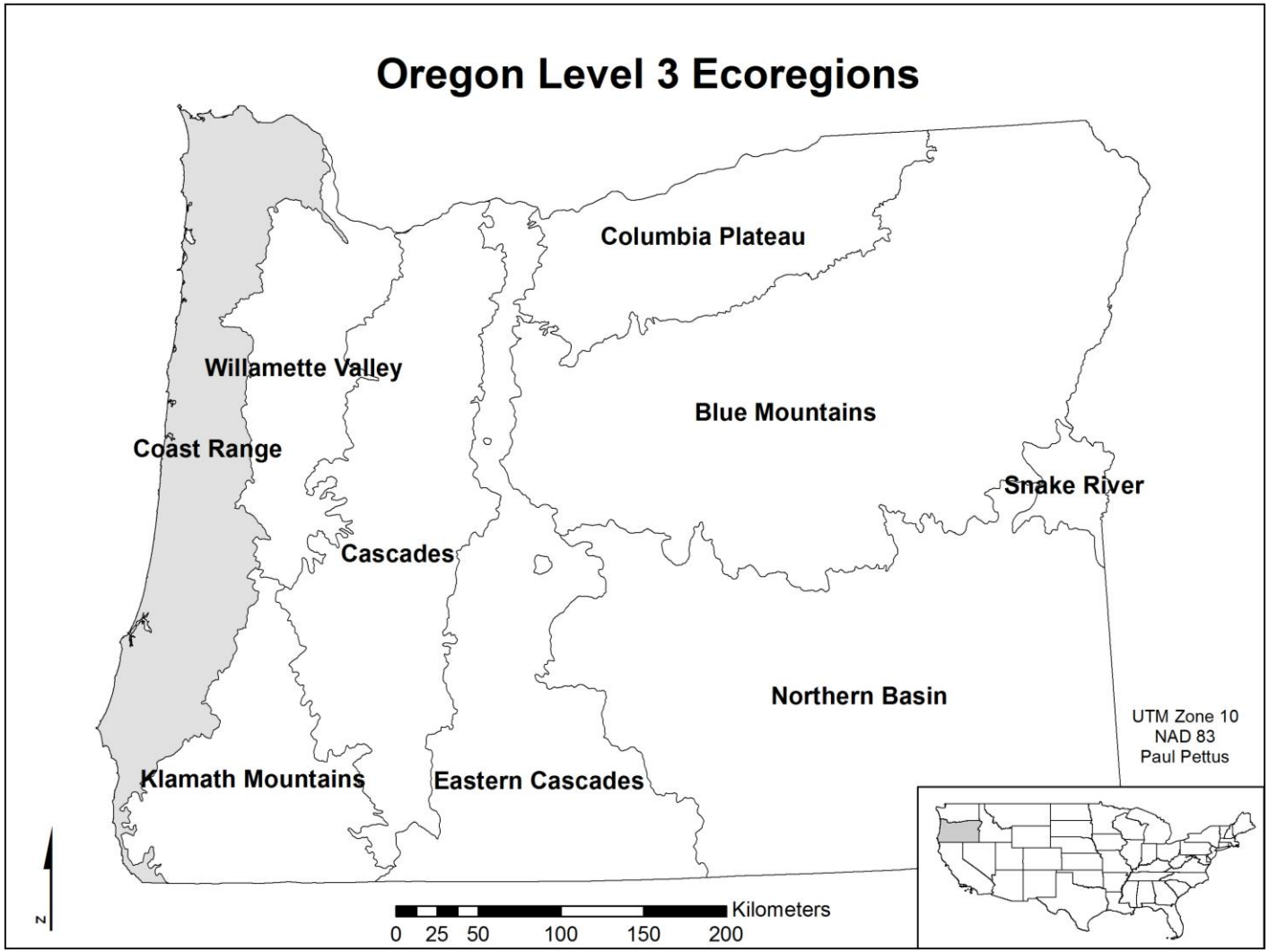


Figure 3. Study area, Oregon Coast Range, USEPA ecoregion level III (Clarke & Schaedel, 1991).

Water Quality Data

Approximately 16,400 fecal coliform and *E. coli* sample counts between the year 2000 and 2010 were collected by ODEQ or partnering organizations. These data along with station sampling location data were retrieved from ODEQ's online Laboratory Analytical Storage and Retrieval (LASAR) database (ODEQ, 2012) (Table 3). Only Quality Control (QC) water quality data of A or A+, the highest standards defined in Oregon's "Quality Assurance Project Plan" were collected for this project (ODEQ, 2008). According to ODEQ protocol when assessing water quality in relation to *E. coli* counts, maximum probability of the number (MPN/100ml) and colony forming units (CFU/100ml) were considered equal, and translated to a generic count number in this analysis (R. Michie ODEQ, personal communication). When a MPN or CFU of either fecal coliform or *E. coli* "Result" column contained characters "est"(estimated count #), "<"(less than count #), or ">"(greater than count #) the following protocol was to apply the equation below:

$$C_R * 0.80 = C_N \quad (1)$$

where C_R equaled the reported count and C_N equaled the new count used for analysis and reporting (R. Michie ODEQ,

personal communication). In 1996, the state of Oregon switched to an *E. coli* indicator pathogen organism standard in fresh and estuarine waters, and a fecal coliform standard for estuarine and marine shellfishing waters standard. With the need to make comparisons in estuarine or other waters, to meet water quality standards, a regression equation was made to facilitate easy transference between fecal coliform counts and *E. coli* indicators (Cude, 2005). Since a disproportional amount of the data set's results were reported as *E. coli* indicators the following regression equation from Cude (2005) was used to transform fecal coliform counts to *E. coli* counts:

$$E.\ coli = 0.531 * (\text{Fecal coliform})^{1.06} \quad (2)$$

with Eq. 1 being applied before Eq. 2. "Cancelled", laboratory duplicates, and other miscellaneous anomalies in the count results were removed entirely from the data set.

Table 3. Research data acquisition and sources. Relationships between land use and other watershed variables that influence water quality violations of *E. coli*.(* = Data, Databases Cited)

Organization	Dataset	Data type	File Format	Scale	Uncertainty*
USGS	National Elevation Dataset	Elevation	Raster	1 arc-second	Z value RMSE = 2.44 m
EPA, USGS	NHDPlus Version 2	Hydrography Dataset	Raster	30m	Based off National Elevation Dataset
EPA, USGS	NHDPlus Version 2	Flow / Catchments	Shapefile	1 :100,000	Based off Elevation Dataset
EPA	Ecoregions of North America Level III	Ecoregion	Shapefile	1:3,000,000	Ecoregion development is ongoin
U.S. Dept. Commerce	2010 U.S. Census	Population	Shapefile	Census Block	~0.01% over count
MRLC	National Land Cover Database	Land Cover/Use	Raster	30 m	78% - 85% accurate
PRISM	Climate mapping system	Climate	Raster	30-arcsec	130 m circular error within 90%
ODEQ	Water quality	Sample location	csv	NA	Unknown, see body text.
ODEQ	Water quality	E. coli / Fecal coliform	csv	NA	ODEQ Quality Control level A or A+
USDA	Livestock Census 2007	Livestock	csv	Zip code	NASS's goal is to count all U.S farms,
USDA	State Soil Geographic data base	Soils	Shapefile	1:250,000	Highly dependent on scale and field

Geographic Data

Only publically available data sets were used in this assessment (Table 3). Flow accumulation and hydrography data ($30m^2$) were acquired from the National Hydrography Dataset Plus Version 2.1 (NHDPlusV2) (USEPA, 2012). Digital elevation models ($30m^2$) from National Elevation Dataset (NED) (Gesch et al., 2009) were provided by U.S. Geological Survey. Land use data was from the year 2006 version of National Land Cover Database (NLCD) ($30m^2$) (Fry et al., 2011). Zip code resolution, livestock and animal operations data were retrieved from the U.S. Department of Agriculture's (USDA) National Agricultural Statistics Service database, which had survey data for dates either ending in the year 2007 or 2008 (USDA-NASS, 2009). Census 2010 USA population data at the census block level were retrieved from ERSI, Inc.'s (2012) free ArcGIS Online Map Services. Soil attributes were retrieved from the STATGO soils database (Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture and U.S. General Soil Map (STATSGO2)). ODEQ LASAR latitude and longitude along with site descriptions were taken at face value, when aligning sampling sites to streams within the geographical information system (GIS) platform. Sampling station location was then

placed on the listed stream reach dictated by the descriptor indicated in "Station Memo" and "Station Description" fields reported in the LASAR database. Because of differences in environmental factors such as dilution, osmotic stress, pH, nutrients, and temperature, station selection was limited to non-estuarine lotic streams that did not occur in marine mixing zones (Rozen & Belkin, 2001).

From the acquired data sets, eighty-eight watershed characteristics were derived to match the common environmental factors affecting the survival and transport of *E. coli* in a watershed (Table 4). Besides individual NLCD land use types, four alternate classes or general land use types were also developed. These four classes were a forest set; urban, natural, which aggregated these individual land cover types. From the USDA livestock census, five sets of confined feeding operations were made: sheep, chickens, cattle, milk-dairy, and total operations per zip code. Soils, livestock, population, and climate data had varying scales of resolution and therefore were converted into grid rasters ($30m^2$) to match hydrography and land cover data. Soils predictors were limited to a likely transport zone, and therefore only derived to a depth of 10 cm for each variable. NHDPlusV2 flowlines were used to make two additional brackets

of predictors. These were meant to represent riparian land use directly next to streams, and were classified by two buffered zones of 30m and 100m outwards of the streams. Within these additional riparian catchments zones, soils/physiography, and the land use classes completed the set of 88 watershed characteristics.

Table 4. Complete list of watershed fecal coliform predictors.

Variable	Model Name	Description / notes
Open Water	LU_11	Open Water
Ice/Snow	LU_12	Ice/Snow
Developed, Open Space	LU_21	Developed, Open Space
Developed, Low Intensity	LU_22	Developed, Low Intensity
Developed, Medium Intensity	LU_23	Developed, Medium Intensity
Developed High Intensity	LU_24	Developed High Intensity
Barren Land (Rock/Sand/Clay)	LU_31	Barren Land (Rock/Sand/Clay)
Deciduous Forest	LU_41	Deciduous Forest
Evergreen Forest	LU_42	Evergreen Forest
Mixed Forest	LU_43	Mixed Forest
Shrub/Scrub	LU_52	Shrub/Scrub
Grassland/Herbaceous	LU_71	Grassland/Herbaceous
Pasture/Hay	LU_81	Pasture/Hay
Cultivated Crops	LU_82	Cultivated Crops
Woody Wetlands	LU_90	Woody Wetlands
Emergent Herbaceous Wet-lands	LU_95	Emergent Herbaceous Wet-lands
Natural	Natural	Natural
Urban	Urban	Sum of: LU_21, LU_22, LU_23, LU_24
Agricultural	Ag	Sum of: LU_81, LU_82
Forest	Forest	Sum of: LU_41, LU_42, LU_43
Elevation	Ele	Meters * 100
Slope	Slope	Degrees

Silt	Silt	Percent silt - Top 10cm
Clay	Clay	Percent clay - Top 10cm
Sand	Sand	Percent Sand - Top 10cm
Ksat	Ksat	Saturated hydraulic conductivity (m/s)
Available water	AW	Volume of water available (mm) - Top 10cm
Human Population	Pop	Count of population
Sheep	Sheep	Sheep operations * 1000
Cattle	Cattle	Cattle operations * 1000
Milk	Milk	Dairy operations * 1000
Chicken	Chicken	Chicken operations * 1000
Total Operations	TO	Total animal operations
Temp Max	Tmax	Mean 1991-2010 maximum temperature C°
Temp Min	Tmin	Mean 1991-2010 minimum temperature C°
Precipitation	Precip	Mean 1991-2010 precipitation
Open Water	LU_11_30m	30 meter stream buffered
Ice/Snow	LU_12_30m	30 meter stream buffered
Developed, Open Space	LU_21_30m	30 meter stream buffered
Developed, Low Intensity	LU_22_30m	30 meter stream buffered
Developed, Medium Intensity	LU_23_30m	30 meter stream buffered
Developed High Intensity	LU_24_30m	30 meter stream buffered
Barren Land (Rock/Sand/Clay)	LU_31_30m	30 meter stream buffered
Deciduous Forest	LU_41_30m	30 meter stream buffered
Evergreen Forest	LU_42_30m	30 meter stream buffered
Mixed Forest	LU_43_30m	30 meter stream buffered
Shrub/Scrub	LU_52_30m	30 meter stream buffered
Grassland/Herbaceous	LU_71_30m	30 meter stream buffered
Pasture/Hay	LU_81_30m	30 meter stream buffered
Cultivated Crops	LU_82_30m	30 meter stream buffered
Woody Wetlands	LU_90_30m	30 meter stream buffered
Emergent Herbaceous Wetlands	LU_95_30m	30 meter stream buffered
Natural	Natural_30m	30 meter stream buffered
Urban	Urban_30m	30 meter stream buffered
Agricultural	Ag_30m	30 meter stream buffered
Forest	Forest_30m	30 meter stream buffered

Slope	Slope_30m	30 meter stream buffered
Silt	Silt_30m	30 meter stream buffered
Clay	Clay_30m	30 meter stream buffered
Sand	Sand_30m	30 meter stream buffered
Ksat	Ksat_30m	30 meter stream buffered
Available water	AW_30m	30 meter stream buffered
Open Water	LU_11_100m	100 meter stream buffered
Ice/Snow	LU_12_100m	100 meter stream buffered
Developed, Open Space	LU_21_100m	100 meter stream buffered
Developed, Low Intensity	LU_22_100m	100 meter stream buffered
Developed, Medium Intensity	LU_23_100m	100 meter stream buffered
Developed High Intensity	LU_24_100m	100 meter stream buffered
Barren Land (Rock/Sand/Clay)	LU_31_100m	100 meter stream buffered
Deciduous Forest	LU_41_100m	100 meter stream buffered
Evergreen Forest	LU_42_100m	100 meter stream buffered
Mixed Forest	LU_43_100m	100 meter stream buffered
Shrub/Scrub	LU_52_100m	100 meter stream buffered
Grassland/Herbaceous	LU_71_100m	100 meter stream buffered
Pasture/Hay	LU_81_100m	100 meter stream buffered
Cultivated Crops	LU_82_100m	100 meter stream buffered
Woody Wetlands	LU_90_100m	100 meter stream buffered
Emergent Herbaceous Wetlands	LU_95_100m	100 meter stream buffered
Natural	Natural_100m	100 meter stream buffered
Urban	Urban_100m	100 meter stream buffered
Agricultural	Ag_100m	100 meter stream buffered
Forest	Forest_100m	100 meter stream buffered
Slope	Slope_100m	100 meter stream buffered
Silt	Silt_100m	100 meter stream buffered
Clay	Clay_100m	100 meter stream buffered
Sand	Sand_100m	100 meter stream buffered
Ksat	Ksat_100m	100 meter stream buffered
Available water	AW_100m	100 meter stream buffered

Geoprocessing and Model Building

Initially, ArcGIS 10.0 Service pack 5 (ESRI, 2012) geographical information system was used to analyze all spatial data for this study. It was possible to generate spatially explicit zonal statistics for each of the watershed variables within the ArcGIS environment, but due to the extreme size of the study area, inefficiency, and exaggerated models times, geoprocessing data in ArcGIS became a common problem. Even when combined with the "ModelBuilder" toolset in ArcGIS and custom Python 2.6 (Python Software Foundation, 2010) scripts, geospatial analytics would frequently overwhelm these tools when aggregating data for 10,000 plus subcatchments. A novel approach of using NHDPlusV2 uniquely identified flow catchments and their flow to and flow from entries in the NHDPlusV2 database. This was used to generate a watershed weighted value of all predictors for each catchment in the study area. Each catchment in the NHDPlusV2 dataset has an identifier and relationship entry in the database that indicates flow direction, and whether it flows into another downstream catchment or not. From these relationships a to:from data dictionary was built for each catchment where one could look up and aggregate all of the contributing catchments for any downstream catchment. With

this it was then possible to weight each catchment by its percentage of contributing land use type or other model predictors. As an example, in Figure 4. NHD Catchment ID 23876079 is a flow through catchment and has a contributing area of many upstream flow through catchments as well as true watershed catchments. So, to account for this and differences in catchment sizes, predictors had to be weighted by their relative contribution areas. This custom approach becomes important when visualizing the final model predictions. For these and other geospatial statistical techniques used in this analysis, custom spatial processing scripts were made using R 2.15.2 statistical package (R Core Team, 2012). These scripts were then combined with Python processing to develop effective ways to compile and analyze these data (Appendix A).

Figure 5 diagrams the process flow used to generate the final, spatially explicit model. Water quality sampling stations in the coastal ecoregion were initially parsed down from the full set of 532 stations to non-estuarine lotic streams that did not occur in marine mixing zones location. Further analysis focused to incorporate general temporal trends in the region, water quality sites were therefore also limited to sites that had at least 20 observations that

generally spanned quarterly sampling over the years 2000-2010. The years from 2000 to 2010 are considered, for Oregon, to be a prime candidate sampling period which includes: drought, wet, cool, and record heat years (H. Lee, US EPA, personal communication). This temporal selection, along with natural log geometric averaging:

$$\ln(\sqrt[n]{X_1 X_{2\ldots} X_n}) \tag{3}$$

limited fluctuations in bacteria observations, and sought to address concerns of temporal autocorrelation of the samples. Spatially, sampling site selection was hindered by clustered measurement locations (Figure 4). Much effort was made to eliminate sites that exhibited drainage nesting and upstream sampling site flowing to another downstream reach to reduce spatial autocorrelation. Additional selection was based in part on equalizing watershed sizes (areas) between sampling locations, and optioning for sites which had a greater number and diversity (temporal) of measurements for the study time span. When obtaining enough sites to sufficiently statistically model was not met, hydrologically nested sites were limited by at least a distance of 5+ kilometers.

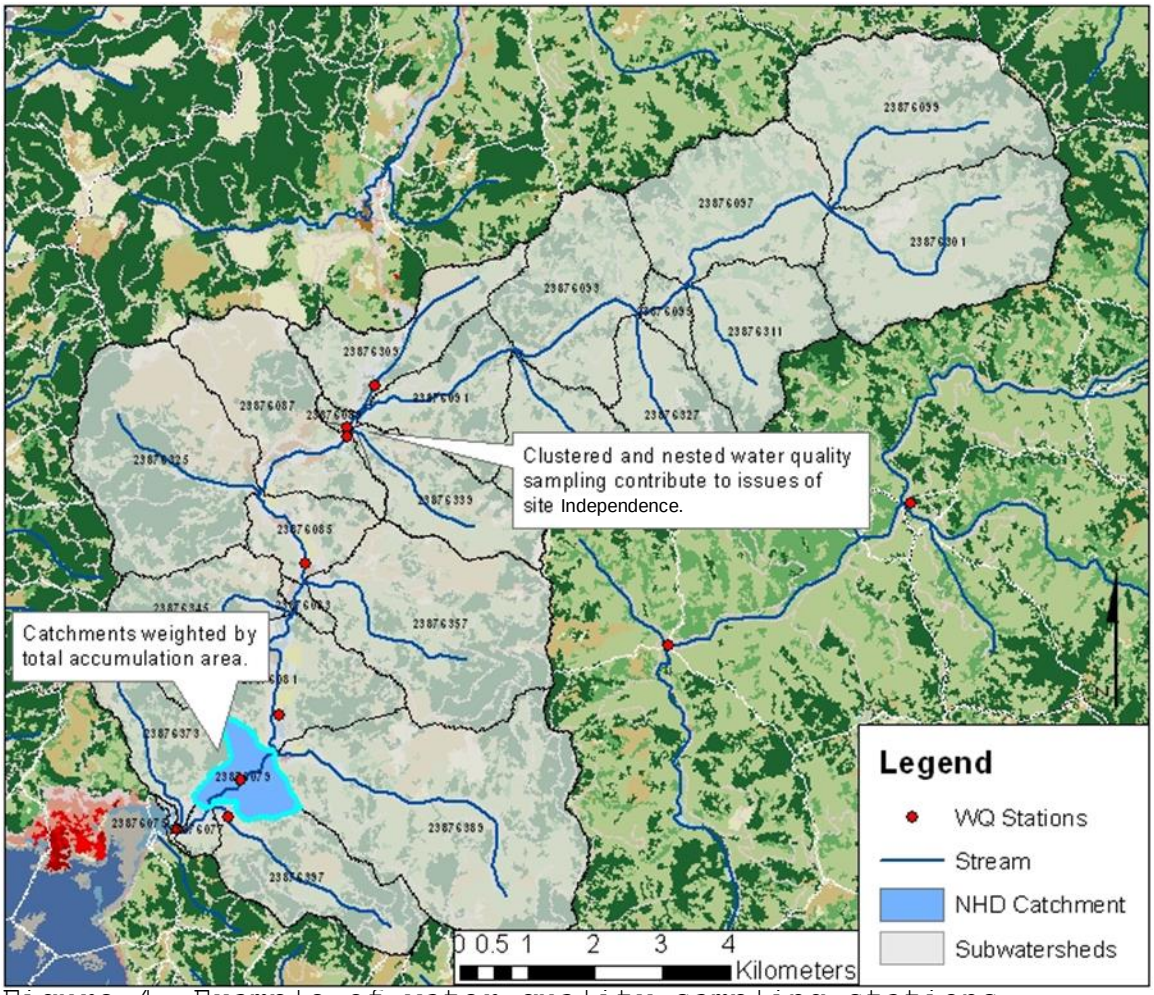


Figure 4. Example of water quality sampling stations, spatial autocorrelation, and site independence issues. Highlighting flow through NHD Catchments and weighting of contributing watershed analysis used in model development and predictions of *E. coli*.

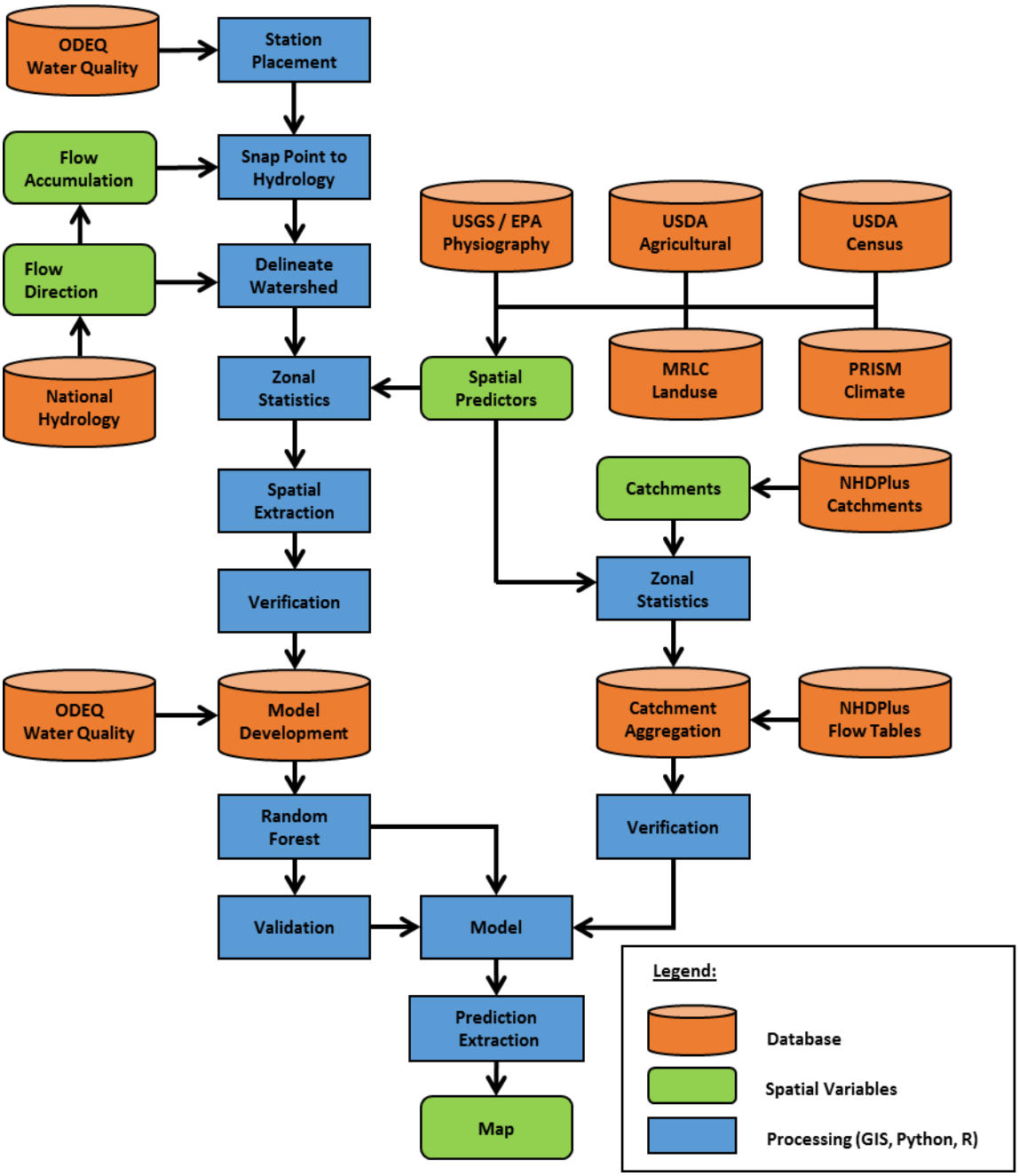


Figure 5. Process flow diagram for spatial analysis of in stream bacteria prediction.

To better understand the general relationships between the predictors themselves, a Principal Components Analysis (PCA) was applied to a subset of the full random forest predictors. Along with animal operations, human population, and general physiography, this PCA was parsed down to just the aggregated riparian buffered (30m and 100m) land use classifications, such as agriculture, forest, urban, and wetlands. PCA, as with multiple regression models, can suffer from over fitting. When too many predictors are added to these models, they can inflate its results. Since the PCA is an exploratory tool, the predictors were reduced to the combined land cover classes, population, and animal operations, from an original 1 site to 1 predictor to a more manageable 1:4 ratio. Interpretable components of the PCA were selected through the broken stick model (Frontier, 1976).

Among other things, multivariate normality of these environmental variables were not fixed by conventional data transformations, so relationships to bacteria could not be explored with many multivariate techniques that require multinormality assumption. Classification trees, however, do not require such assumptions, and can successfully deal with missing data points, non-normality, and unequal variances (Strobl et al, 2009; Torsten et al, 2010). Classification

trees build upon binary splits in the predictor variables to classify a categorical dependent variable. The final prediction model used was a random forest model and is analogues to an ensemble of classification trees. The random forest model was built using a continuous response variable. This non-parametric approach was done using the "randomForest" package in the statistical software R. The random forest modeled spatially explicit watershed variables vs. continuous observations of *E. coli* (Appendix B). This implementation of Breiman's random forest (randomForest) fixed problems that it had towards highly correlated variables (Strobl et al, 2008). A total of 10,000 trees were grown. Variables are said to be important predictors if their variable importance score is higher than the absolute value of the lowest predictor (Strobl et al, 2009). The rationale for this importance of predictors is that "irrelevant variables vary randomly around zero" (Strobl et al, 2009). For visualization purposes and to have a reasonable estimation of what is happening in the random forest model a single Classification and Regression Tree (CART) was also grown. *E. coli* was grouped into three almost equally sized categories: 0-25 (cfu/100ml), 25-50 (cfu/100ml), and 50+ (cfu/100ml) for the CART model. The CART model employed to

expose the complex interaction between the numerous predictor variables, and to give a visual sense of what was likely going on in the random forest model. Finally, the random forest model, along with the flow:to flow:from NHDPlus V2 catchments, a catchment area weighted prediction map was developed for the Oregon coast range. With, R 2.15.2 statistical software (R Core Team, 2012) being used for all analysis, and packages randomForest and rpart for the random forest and CART models.

RESULTS

Of the coast range's 532 sampling locations retrieved from ODEQ's online database a final study set of 93 sites was compiled. These sites were chosen due to reasons of: salinity in tidal zones, watershed nesting, station sampling counts, and temporal diversity among other things (Figure 6). More broadly, this selection left a more northern grouping of sites than in the southern coast range, with approximately two thirds of the sites being to the north of the city Newport, a gap of few sites in the central coast, and other third spread along the southern region. These sites, in total had 6657 samples collected during the study years (2000-2010), averaging roughly 70 samples per site. Land use between the watersheds varied considerably: Agriculture 0% - 7%, Forest 48% - 91%, Urban 2% - 11%, and Natural 85% - 98%, with means of 1%, 70%, 6%, and 93% respectively (Table 5). Study watershed size ranged from a 25% quartile of ~4,800 ha to 75% quartile of ~42,500 ha, and a mean of 32,300 ha. Geomean *E. coli* counts ranged from 5 (cfu/100ml) to 396 (cfu/100ml) with a median of 36 (cfu/100ml).

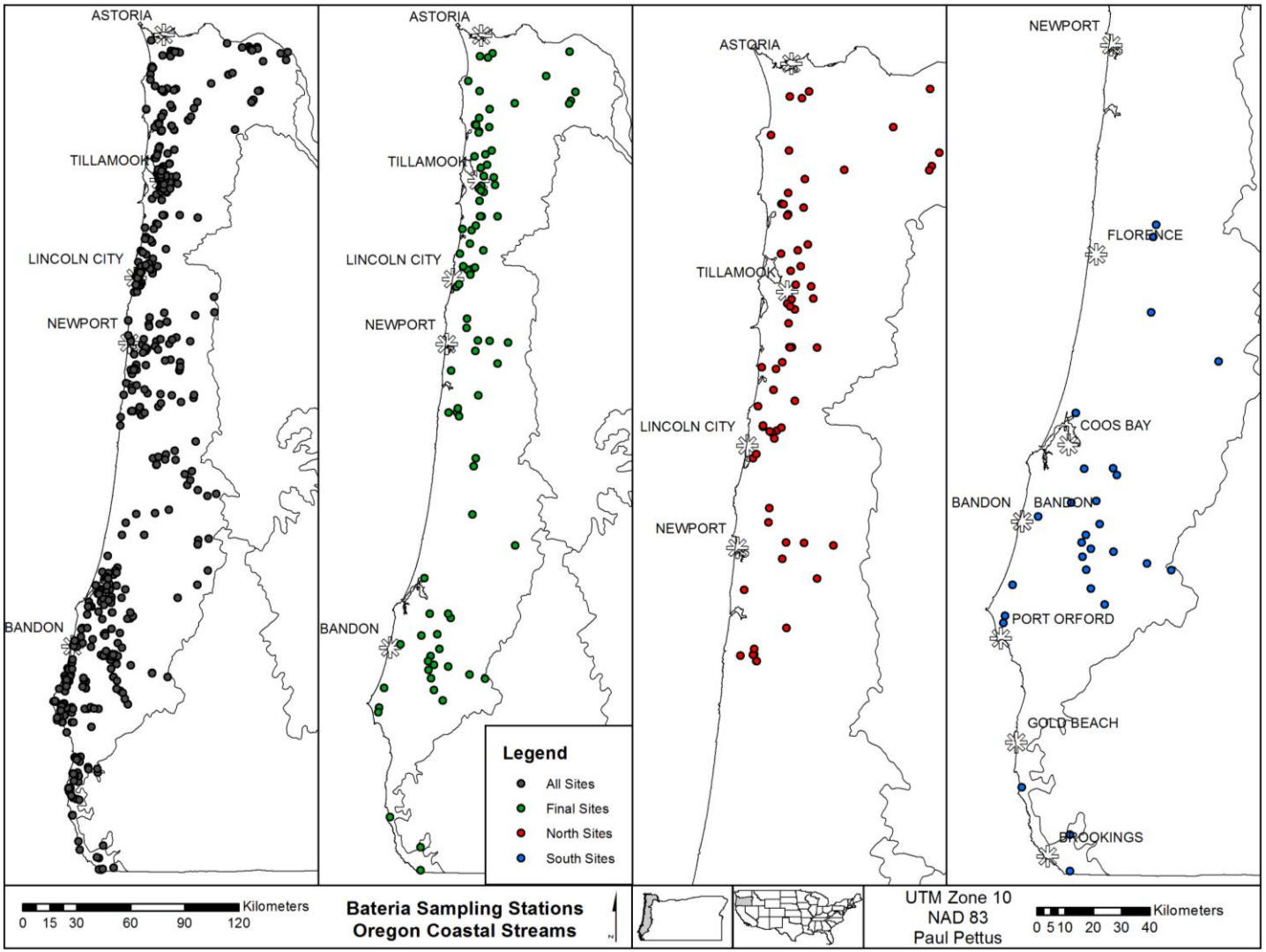


Figure 6. Oregon coast range ecoregion bacteria sampling stations (left), final selection (center left), north coast sites (center right), and south coast sites (right).

Table 5. Summary statistics of final study watersheds, predictors, and fecal coliform. (Q. = Quartile) (Units: Population, and animal operations are average #/30m². All land uses and soils are in % of watershed. Slope is average # of degrees (slope angle). Elevation is cm)

Variable	Min	1st Q.	Median	Mean	3rd Q.	Max
Watersheds (ha)	303	4880	14700	32300	42500	191000
Ecoli_geomean (cfu/100ml)	4.5	20.6	33.6	60.4	67.5	396.0
awc	0.16	0.19	0.22	0.23	0.27	0.30
awc_100m	0.17	0.19	0.21	0.23	0.27	0.30
awc_30m	0.17	0.19	0.21	0.23	0.27	0.30
cattle	0.0027	0.0303	0.0665	0.0648	0.0914	0.1690
chick	0.0011	0.0091	0.0138	0.0173	0.0197	0.0733
clay	17.80	20.20	21.00	21.50	22.50	30.90
clay_30m	18.40	20.50	21.00	21.70	22.50	35.50
clay_100m	18.40	20.50	21.00	21.70	22.50	35.20
elevation	8130	24300	29300	32600	38500	72100
ksat	6.48	9.17	9.17	14.10	20.60	28.20
ksat_30m	6.15	9.17	9.17	13.90	19.40	28.20
ksat_100m	6.15	9.17	9.17	13.90	19.50	28.20
milk	0.0000	0.0019	0.0035	0.0119	0.0133	0.0619
population	0.0000	0.0008	0.0017	0.0041	0.0033	0.0904
precip	116000	198000	241000	245000	295000	378000
sand	7.20	15.40	23.50	24.00	32.90	42.10
sand_30m	7.20	15.30	21.80	23.40	32.20	42.10
sand_100m	7.20	15.40	21.70	23.50	32.20	42.10
sheep	0.00	0.00	0.01	0.01	0.02	0.07
silt	37.90	44.50	55.60	55.00	63.60	70.30
silt_30m	37.90	46.50	56.60	55.20	63.40	70.30
silt_100m	37.90	46.30	56.40	55.20	63.30	70.30
slope	9.89	14.80	17.10	17.70	19.20	31.80
slope_30m	5.80	8.72	11.90	11.70	14.20	18.90
slope_100m	8.10	12.80	15.40	15.60	17.90	24.80
temp_max	1360	1470	1520	1550	1640	1790
temp_min	423	517	555	554	601	685
forest	0.4860	0.6140	0.7040	0.7060	0.7830	0.9110
ag	0.0000	0.0000	0.0043	0.0107	0.0142	0.0692
natural	0.8520	0.9240	0.9340	0.9330	0.9480	0.9830

urban	0.0164	0.0459	0.0509	0.0562	0.0684	0.1100
LU_21	0.0161	0.0416	0.0503	0.0536	0.0666	0.1020
LU_41	0.0004	0.0094	0.0186	0.0253	0.0333	0.1290
LU_42	0.2000	0.4000	0.4460	0.4620	0.5180	0.7670
LU_43	0.0238	0.1540	0.2100	0.2190	0.2750	0.4890
LU_52	0.0300	0.1010	0.1380	0.1550	0.2030	0.3570
LU_71	0.0000	0.0291	0.0572	0.0617	0.0822	0.2030
LU_90	0.0000	0.0038	0.0058	0.0066	0.0082	0.0286
LU_11	0.0000	0.0000	0.0000	0.0002	0.0002	0.0025
LU_22	0.0000	0.0003	0.0013	0.0023	0.0027	0.0355
LU_23	0.0000	0.0000	0.0001	0.0002	0.0002	0.0041
LU_24	0.0000	0.0000	0.0000	0.0001	0.0001	0.0006
LU_31	0.0000	0.0005	0.0011	0.0016	0.0020	0.0090
LU_81	0.0000	0.0000	0.0039	0.0101	0.0132	0.0685
LU_82	0.0000	0.0000	0.0001	0.0006	0.0007	0.0062
LU_95	0.0000	0.0004	0.0009	0.0020	0.0019	0.0276
ag_100m	0.0000	0.0000	0.0090	0.0236	0.0270	0.2820
forest_100m	0.4050	0.6010	0.6740	0.6790	0.7650	0.9920
urban_100m	0.0024	0.0663	0.0784	0.0825	0.0927	0.1890
natural_100m	0.6730	0.8750	0.9040	0.8940	0.9200	0.9980
LU_100m_21	0.0024	0.0645	0.0755	0.0782	0.0927	0.1890
LU_100m_41	0.0003	0.0136	0.0327	0.0472	0.0577	0.3240
LU_100m_42	0.0717	0.2260	0.2980	0.3030	0.3520	0.6410
LU_100m_43	0.0566	0.2690	0.3320	0.3290	0.3910	0.5580
LU_100m_52	0.0031	0.0766	0.1050	0.1140	0.1430	0.2830
LU_100m_71	0.0000	0.0242	0.0433	0.0534	0.0758	0.1970
LU_100m_90	0.0000	0.0223	0.0325	0.0375	0.0484	0.1450
LU_100m_11	0.0000	0.0000	0.0000	0.0011	0.0012	0.0115
LU_100m_22	0.0000	0.0002	0.0014	0.0038	0.0042	0.0643
LU_100m_23	0.0000	0.0000	0.0000	0.0004	0.0003	0.0088
LU_100m_24	0.0000	0.0000	0.0000	0.0001	0.0000	0.0009
LU_100m_31	0.0000	0.0002	0.0010	0.0013	0.0023	0.0058
LU_100m_81	0.0000	0.0000	0.0086	0.0221	0.0246	0.2820
LU_100m_82	0.0000	0.0000	0.0000	0.0015	0.0010	0.0263
LU_100m_95	0.0000	0.0013	0.0037	0.0081	0.0075	0.1160
ag_30m	0.0000	0.0000	0.0061	0.0211	0.0235	0.2990
urban_30	0.0000	0.0291	0.0354	0.0425	0.0555	0.1210
natural_30m	0.6780	0.9250	0.9490	0.9360	0.9630	1.0000
forest_30m	0.3550	0.6510	0.7220	0.7150	0.8070	1.0000

LU_30m_21	0.0000	0.0283	0.0344	0.0398	0.0531	0.1160
LU_30m_42	0.0139	0.1790	0.2340	0.2560	0.3210	0.6390
LU_30m_43	0.0820	0.3310	0.4060	0.4010	0.4760	0.6880
LU_30m_52	0.0000	0.0584	0.0900	0.0987	0.1340	0.2730
LU_30m_71	0.0000	0.0218	0.0297	0.0450	0.0598	0.2370
LU_30m_90	0.0000	0.0371	0.0521	0.0648	0.0847	0.2820
LU_30m_41	0.0000	0.0158	0.0388	0.0572	0.0718	0.3500
LU_30m_11	0.0000	0.0000	0.0000	0.0017	0.0014	0.0164
LU_30m_22	0.0000	0.0000	0.0007	0.0025	0.0023	0.0610
LU_30m_31	0.0000	0.0000	0.0006	0.0010	0.0013	0.0087
LU_30m_81	0.0000	0.0000	0.0048	0.0199	0.0200	0.2990
LU_30m_82	0.0000	0.0000	0.0000	0.0012	0.0007	0.0297
LU_30m_95	0.0000	0.0018	0.0048	0.0105	0.0106	0.1690
LU_30m_23	0.0000	0.0000	0.0000	0.0002	0.0001	0.0042
LU_30m_24	0.0000	0.0000	0.0000	0.0000	0.0000	0.0005
T_operations	0.0042	0.0531	0.1110	0.1080	0.1490	0.2460

Variables for the PCA were trimmed down to a final 21 predictors, for a ratio of nearly four sites to for each predictor. Spatially, sites were placed into one of two categories, north and south, based roughly on a half-way point in the coastal region and the visual patterns seen in the data (Figure 6). Through the broken-stick model the PCA was reduced to 4 principal components explaining a total variance of 74% (Table 6). In the first principal component (PC) bank slope, forested, and natural land uses most strongly and positively correlated together, while the variables related to agriculture and wetlands had nearly as strong negative correlations (Table 7, Figure 7). Within the second component, grasslands had the highest positive loading, and

urban land use and sheep operations loaded negatively. Lastly in the third and fourth components, animal operations variables had the strongest negative and positive loadings respectively. In the Figure 7 it is apparent that natural riparian zone and agricultural areas have opposite vectors in PC one and urban and animal operations become visually negatively orthogonal on the second PC (Figure 7). North and south locations appear to randomly spread over both PC one and PC two.

Table 6. Total variance explained from PCA on broken-stick reduced components.

Component	Eigenvalues		
	Total	% Variance	Cumulative %
1	8.45	0.40	0.40
2	2.81	0.13	0.54
3	2.33	0.11	0.65
4	2.06	0.10	0.74

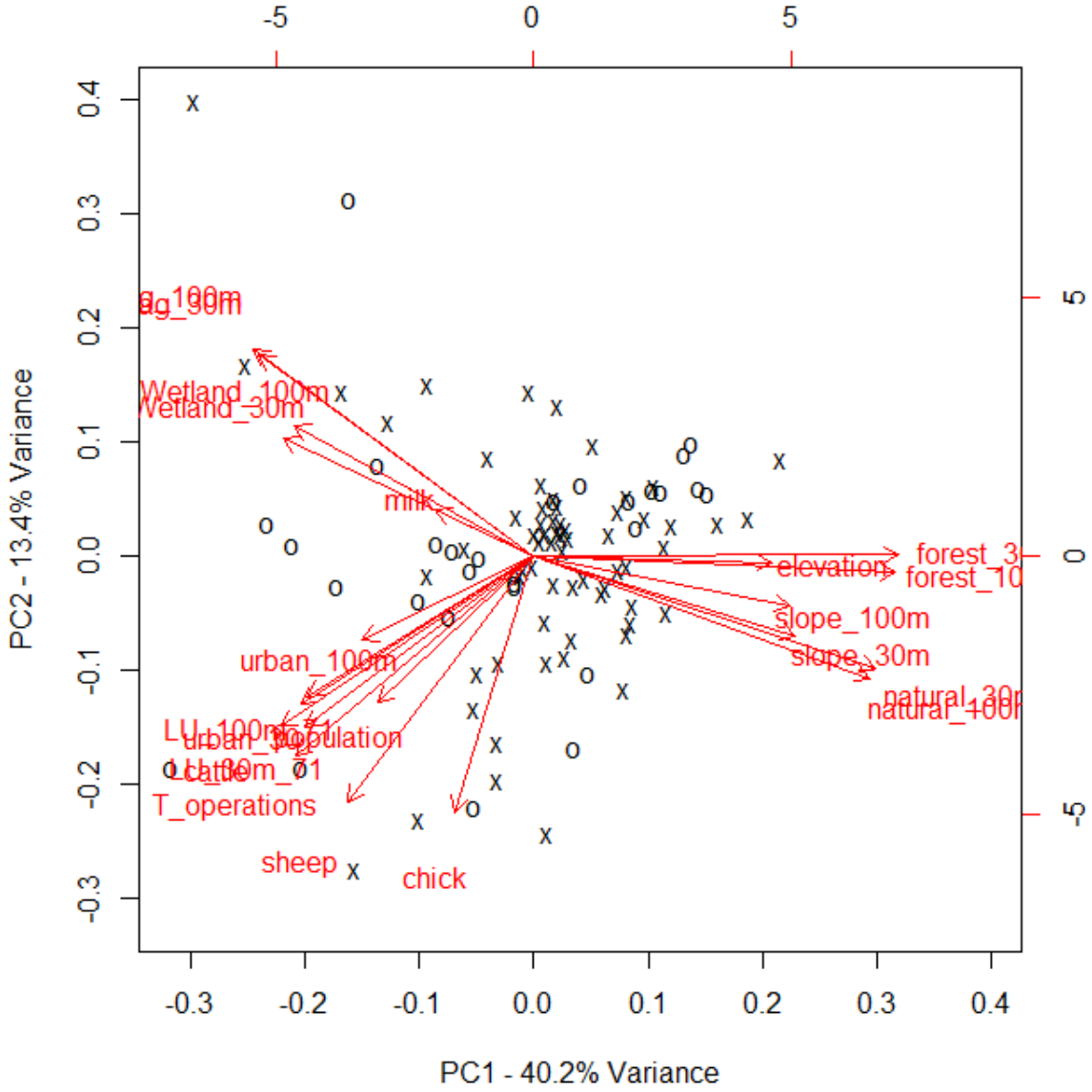


Figure 7. Principal components analysis for reduced watershed predictors (only the first 2 PCA axes were plotted). 0 = South Coast sites, and X = North coastal sites.

Table 7. Eigenvectors, loading for each of the final PCA components. Values were highlighted to call attention to the most influential loadings.

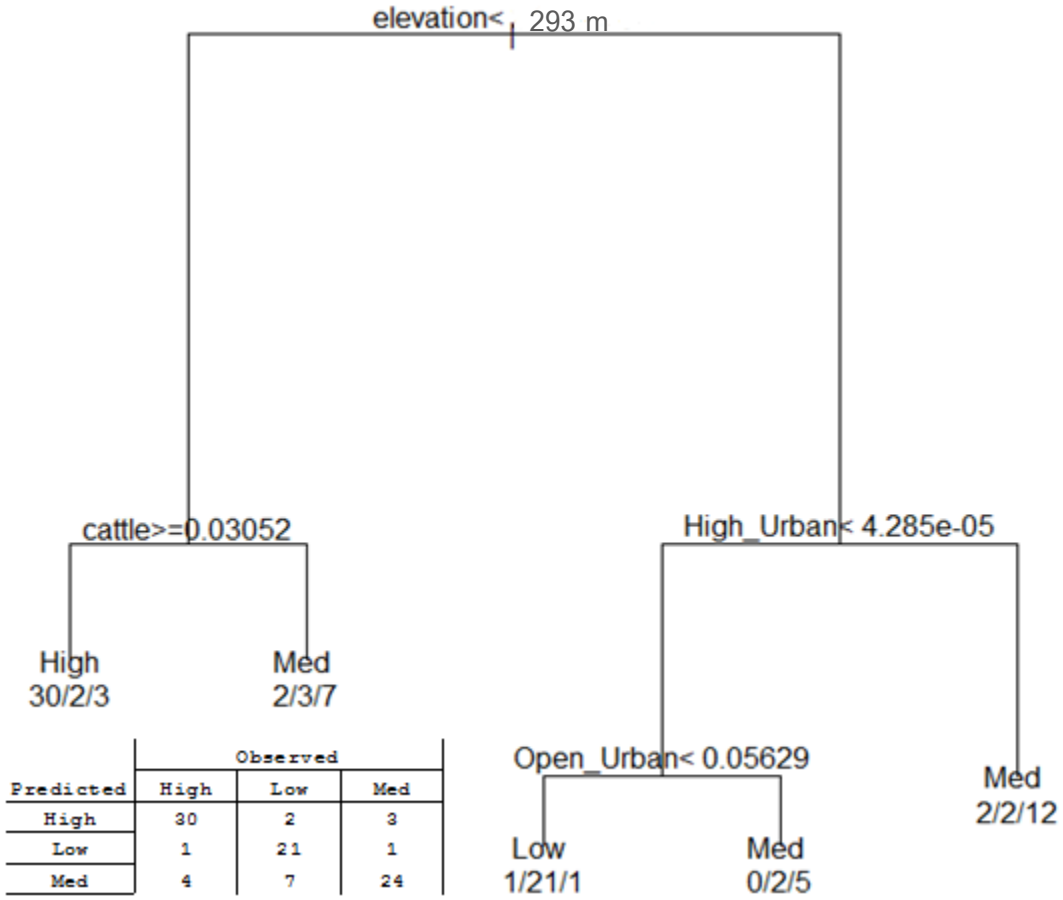
	Principal Component			
Variable	PC1	PC2	PC3	PC4
cattle	-0.19	0.15	-0.12	0.39
chick	-0.05	-0.06	-0.32	0.31
elevation	0.17	0.15	0.06	0.02
milk	-0.08	-0.19	0.08	0.39
population	-0.11	-0.16	-0.27	0.06
sheep	-0.13	0.27	-0.24	0.18
slope_30m	0.21	0.15	0.03	0.13
slope_100m	0.21	0.15	0.08	0.16
forest	0.21	-0.20	0.07	0.14
ag	-0.21	0.05	0.30	0.15
natural	0.22	0.27	-0.02	-0.03
urban	-0.09	-0.38	-0.23	-0.08
LU_71	-0.15	0.33	-0.10	-0.21
ag_100m	-0.22	0.04	0.33	0.14
forest_100m	0.28	-0.12	0.01	0.12
urban_100m	-0.13	-0.29	-0.28	-0.10
natural_100m	0.26	0.15	-0.10	-0.06
LU_100m_71	-0.17	0.36	-0.16	-0.17
ag_30m	-0.22	0.05	0.32	0.14
urban_30	-0.17	-0.12	-0.32	-0.09
natural_30m	0.26	0.01	-0.14	-0.09
forest_30m	0.28	-0.13	0.01	0.09
LU_30m_71	-0.17	0.31	-0.19	-0.10
T_operations	-0.17	0.08	-0.18	0.46
Wetland	-0.21	-0.12	0.14	-0.13
Wetland_30m	-0.20	-0.09	0.14	-0.17
Wetland_100m	-0.19	-0.10	0.15	-0.21

The CART model clearly shows that elevation is the most important factor in prediction of stream fecal coliforms, as it was the primary split of the model (Figure 8). In the lower elevation sites, cattle operations in a watershed were associated with high bacteria counts. While in the higher elevation sites, high intensity development land use was related with primarily medium concentrations of *E. coli.* With areas of lower intensity urban development uses, bacteria counts were predicted to be classified into the low or medium category. The CART model had a 19.4% misclassification rate. The complete predictor random forest model explained %56.5 of the variation, with a Mean of squared residuals of 0.36. The highest values in variance importance plot for the random forest model are primarily giving preference to the combined natural and forested riparian (30m and 100m) land use predictors (Figure 9). Cattle and total animal operations are in the mid to higher range of variable importance. Similarly to the CART model, it also shows that watershed mean elevation as the primary predictor, yet it also yields riparian slope as of high importance. As the other variables importance values near zero, they become relatively unimportant to the random forest model.

A visualization of the Oregon coast range's predicted

catchment level *E. coli* concentration can be examined in Figures 10 – 13. These figures move down the coastal region from north, central, and to the south highlighting the ecoregion's potential for bacteria impairment. In Figure 10, catchments predicted to have higher levels of *E. coli counts* (left panel), such as those close into Tillamook, are also associated with areas of higher agricultural and urban land uses (right panel). Moving down to the central and southern coast, similar mid and high level bacteria prediction follow pastures and urban land use patterns, while higher elevation, forested, and natural areas inland are linked to lower concentration count predictions (Figures 11 - 13). For further reference, and to "ground truth" the accuracy of the random forest model prediction catchments, sampling site locations for all of ODEQ's coastal bacteria stations were also included in the left panels while sites used to build the model are seen in the right panels (Figure 6, 10-13). The reader needs to be aware that these 532 sites are averages of all counts (cfu/100ml) between the study years 2000-2010, and can range from as little as one sample to hundreds of samples per site. For visualization purposes, color coding for ODEQ sampling sites and prediction catchments were standardized through Figures 10-13. As an example, in Figures

12 & 13, from Bandon to the north and east of Coos Bay, the prediction maps fills in unsampled drainages in a similar nature to the sampled streams, and clear relations between land use types, watershed characteristics and bacteria sampling counts in relation to land use types can be seen when comparing between the panels in Figures 10-13.



Predicted	Observed		
	High	Low	Med
High	30	2	3
Low	1	21	1
Med	4	7	24

Figure 8. Classification and regression tree model of in stream *E. coli* for Oregon's coastal streams. 19.4% misclassification rate.

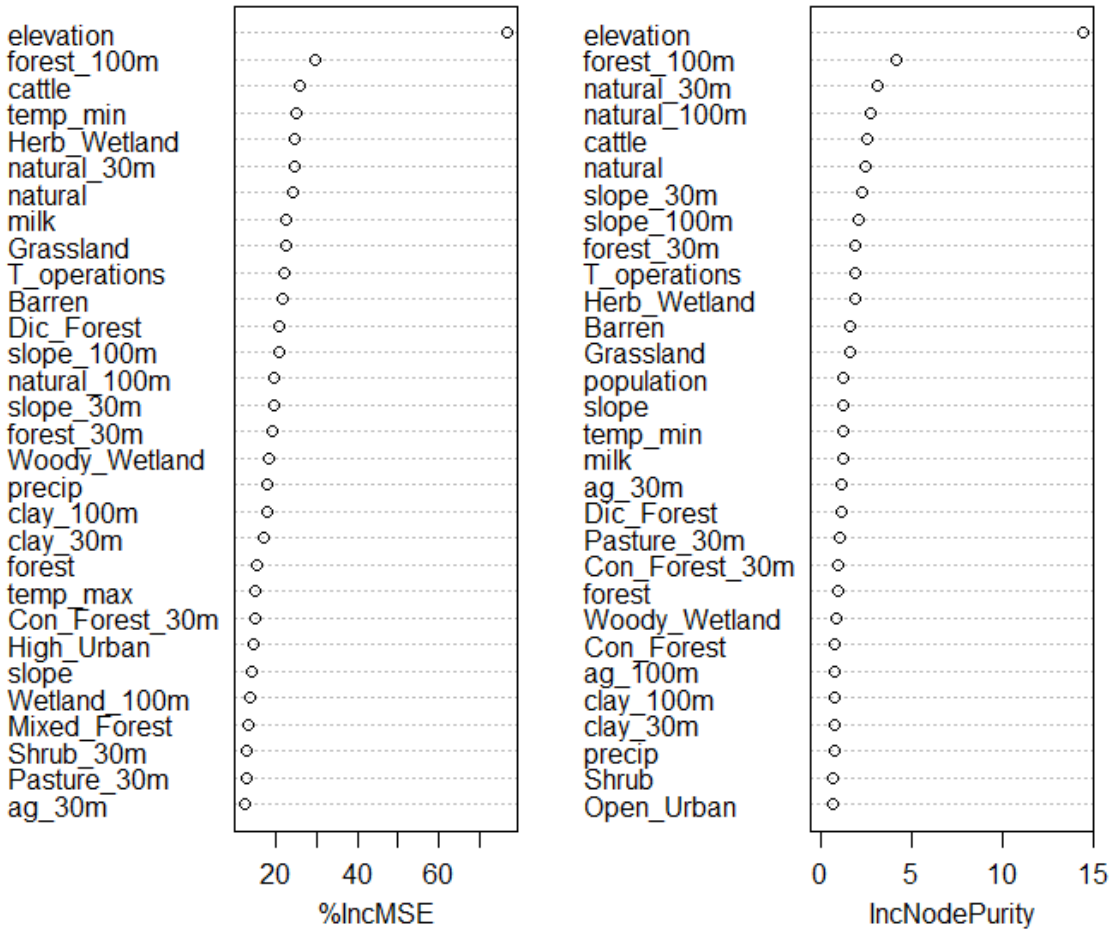


Figure 9. Random forest variable importance plot. Higher variable importance increase node splitting purity, variables closest to zero are relatively unimportant (IncNodePurity). While variables with higher "%IncMSE" increase node impurity when randomly permutated.

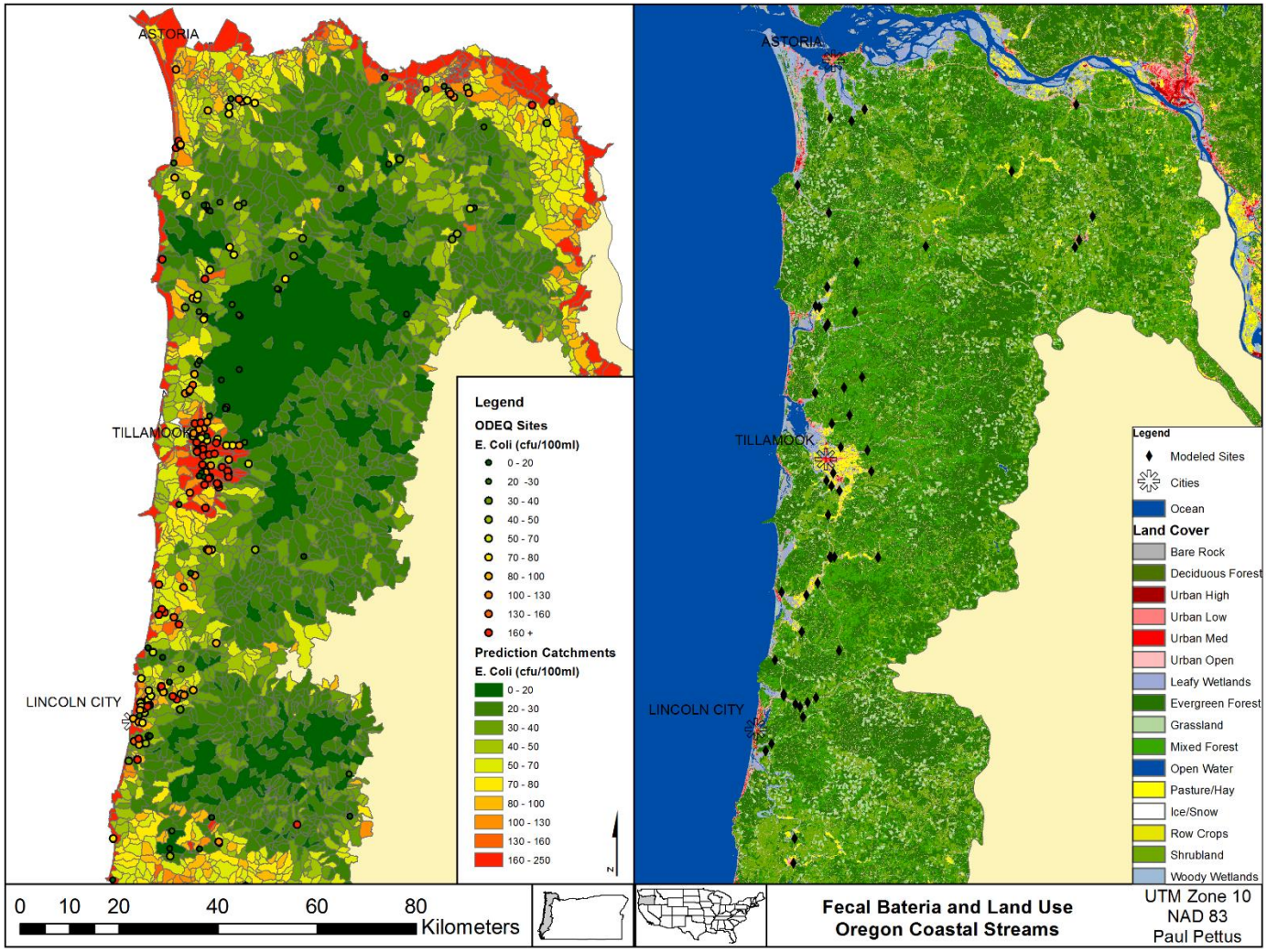


Figure 10. The Left panel includes *E. coli* (CFU /100 ml) predictions of 2000-2010 average in North Oregon coast range stream NHD Catchments, and ODEQ sampling sites averaged counts for the study years. The right panel displays random forest modeled sites and 2006 NLCD land use classifications.

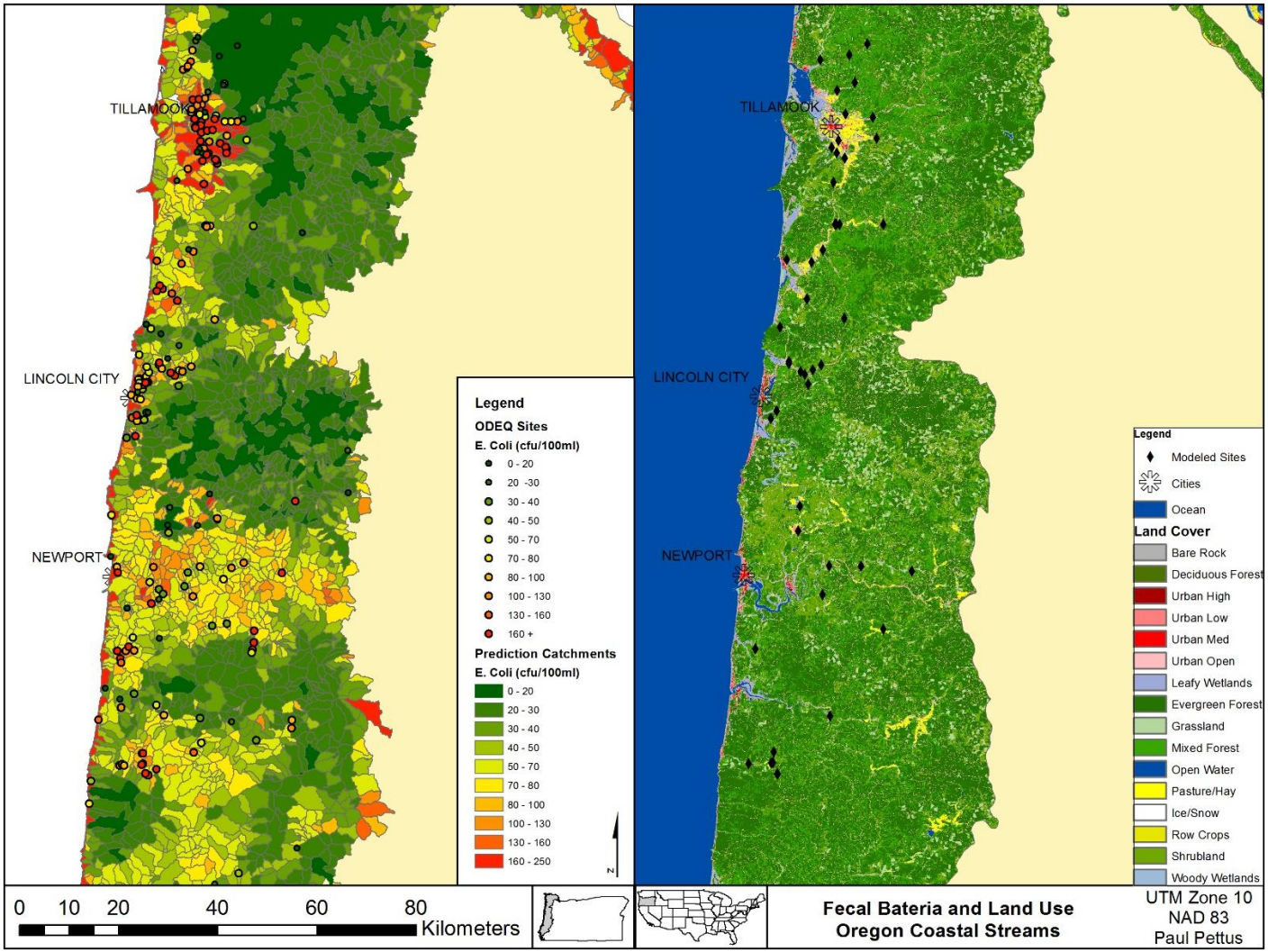


Figure 11. The Left panel includes *E. coli* (CFU /100 ml) predictions of 2000-2010 average in North Central Oregon coast range stream NHD Catchments, and ODEQ sampling sites averaged counts for the study years. The right panel displays random forest modeled sites and 2006 NLCD land use classifications.

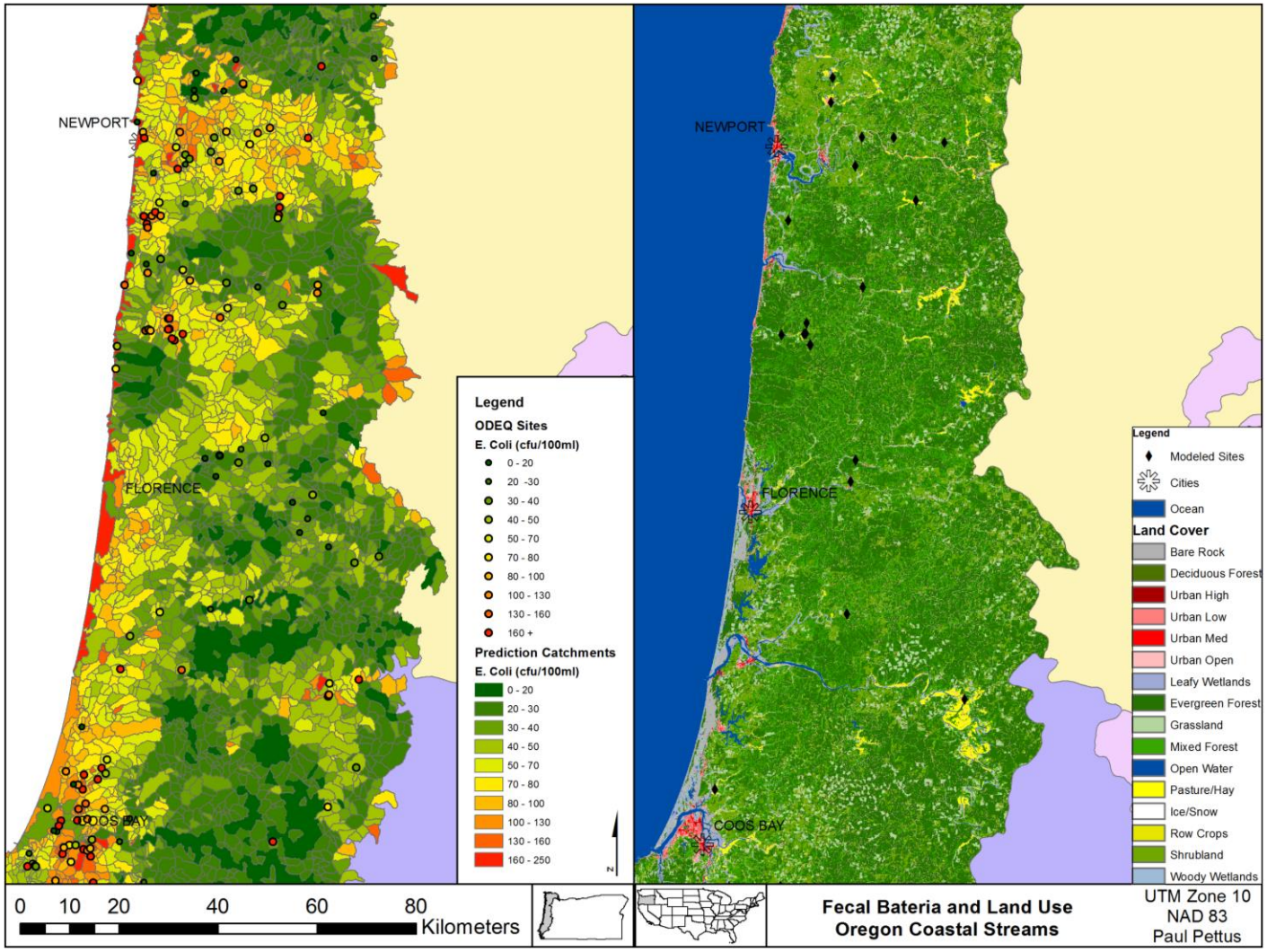


Figure 12. The Left panel includes *E. coli* (CFU /100 ml) predictions of 2000-2010 average in South Central Oregon coast range stream NHD Catchments, and ODEQ sampling sites averaged counts for the study years. The right panel displays random forest modeled sites and 2006 NLCD land use classifications.

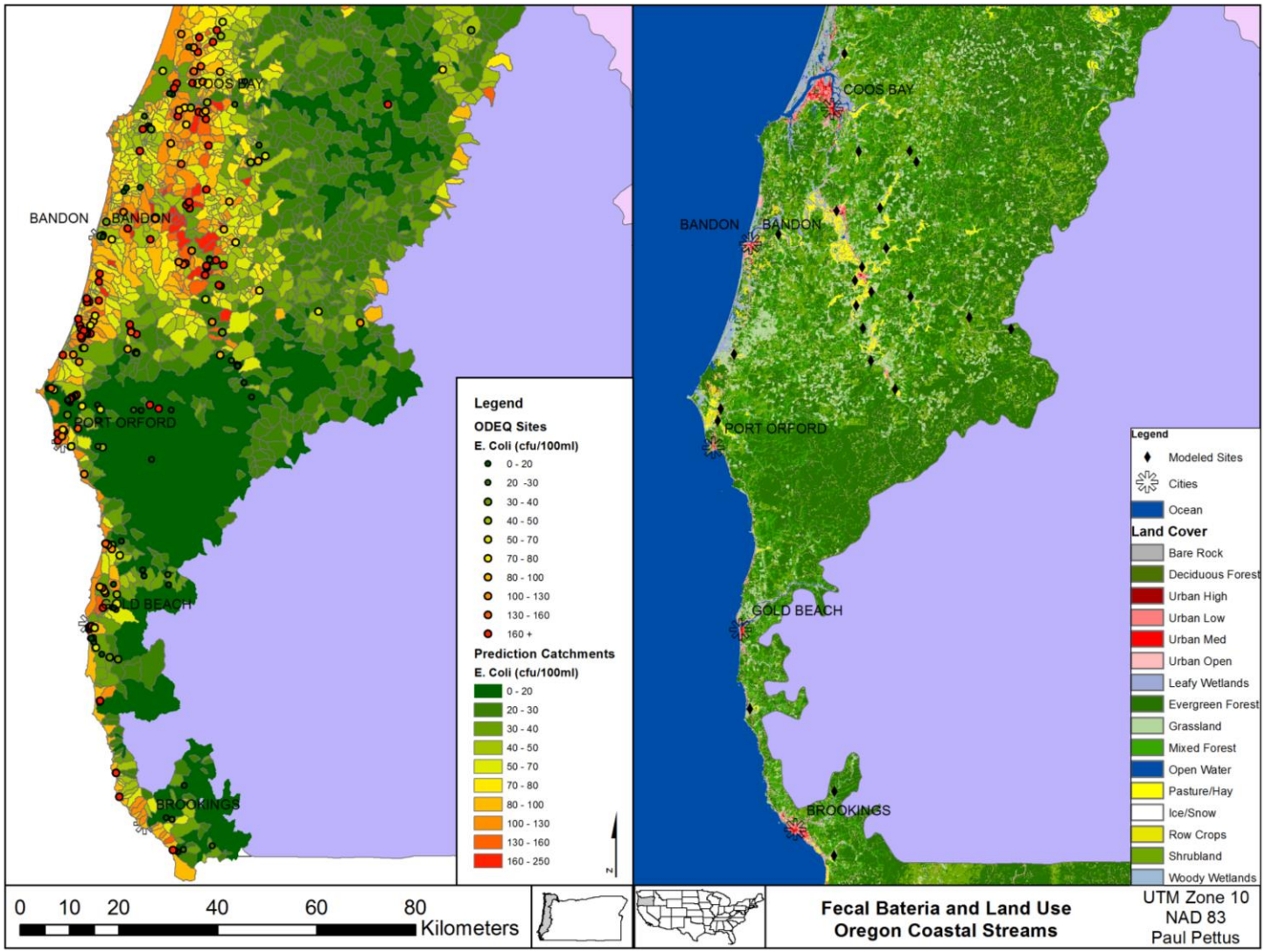


Figure 13. The Left panel includes *E. coli* (CFU /100 ml) predictions of 2000-2010 average in South Oregon coast range stream NHD Catchments, and ODEQ sampling sites averaged counts for the study years. The right panel displays random forest modeled sites and 2006 NLCD land use classifications.

DISCUSSION

The purpose of this research was three fold: to generate a generalized stream bacteria prediction model from easily obtainable watershed characteristics, to identify likely areas of high pathogen bacteria concentrations with infrequent monitoring, and to allow for future land use scenario analysis. This random forest model in essence, provides a 2000-2010 year average, spatial snapshot of likely *E. coli* concentrations throughout Oregon's coastal region. The findings of the random forest model appear sufficient and reliable, when compared to other researches. This model's 56.5 % explanation of variation, analogues to an uninflated R^2 in a regression model, matches with Crowther et al. (2011) stream fecal coliform research on land cover and population related variables. Land use was broken down into four categories of woodland, urban, grassland, and arable, while populations were defined by human, dairy, cattle, and sheep densities. These researchers' regional models had prediction adjusted R^2 values ranging from 0.54 to 0.62 for in-stream fecal coliforms. This United Kingdom study was limited to 14 coastal draining catchments, sampled in the summer bathing season, between the years 1995-2005, and used a minimum of 5 samples for each site under base and high flow conditions to

make their fecal indicator model. While similar research by the Ministry of Environment in New Zealand, researchers' nationwide random forest model could explain 69.8% variation of in-stream *E. coli* (NZME, 2010). These researchers had roughly 400 sampling sites with 5 years of consecutive quarterly sampled bacteria data, and did not make a distinction between independent and nested drainages in their analysis. The majority of these sampling sites were either clustered around the population centers of the North Island, or on the southern portion of South Island. Dissimilarities between this study's random forest model and the Ministry of Environment researchers could be linked to a roughly four fold more sampling sites, precise quarterly sampling, or the clustered sampling locations in New Zealand. Differences in uncertainties in GIS layers could also contribute to a higher explanation of variance in the NZME random forest model. For example, this study used the NLCD 2006 land-cover dataset which had an accuracy of 78% for 16 land use classes, while the NZME used a 43 class Land Cover Database (LCDB) that has a ~96% accuracy rate (LCDB, 2012; Wickham et al. 2013).

Other statistical methods such as the CART analysis both showed linkages to agricultural related activities and urban land uses as being highly influential on bacteria counts

(Figure 7 & 8, Table 7). The PCA showed that predictors related to anthropogenic activities, such as grazing and urban land use, were highly correlated together. These results coincide with Tillamook Bay research on genetic identification and source characterization of fecal pollution (Bernhard et al., 2002). Their research found most fecal coliforms showed genetic markers from dairy operations and sewage due to anthropogenic activities on the coast. The CART and random forest models showed elevation as a primary predictor, which agrees with conventional knowledge, that as one rises into a drainage basin and away from human activity water quality will improve. The random forest model highlighted the importance of riparian land uses over overall watershed land uses. This agrees with the body of evidence showing that natural and/or forest riparian buffers contribute significantly to improvements in water quality (Osborne & Kovacic, 1993; Lowrance et al., 1997). Now that the random forest model has been developed for current regional conditions, future scenarios relating to changes or improvements in riparian zones could be explored.

To assess uncertainties in the model we must first start with the underlying GIS layers. As documented in the metadata of the publicly available datasets, these layers have

reasonable ranges of errors. Again, the National Land Cover Dataset notes a 78% – 85% classification accuracy rate, because it is derived in part from statistical regressions of diverse remote sensing techniques (Fry et al., 2011). Another example is the soils STATSGO2 data which is derived from coarse soil surveys (1:250,000). There are new higher resolution USDA soils data, Soil Survey Geographic (SSURGO) dataset, scaling from 1:12,000 to 1:63,360, but these data have numerous voids on National lands, and could not be used in this analysis. Populations data such as the USDA animal operations were of poor resolution, zip code level, and transformed to counts which were then spatially averaged and assigned equally over a zip code. Human census counts were of finer resolution, because census boundary size is based on population densities. A single census district could be as small as an apartment building which had a population of 200+, or could be expansive, because a rural area might have almost no human residences. Again, sampling site placement was taken at face value from the site descriptors and accuracy of GPS locations. Much care was made to control for spatial and temporal bias, in averaging site samples across a climatically dynamic time span, and to eliminate hydrological connected sites. Yet, infrequently, some site nesting

remained. Additionally the data were not vigorously explored for seasonal diverseness.

There are several possibilities for model improvements and future assessment. ODEQ takes a "Watershed Approach" to define related waterways, and groupings of basins into regions that are similar in geography and to facilitate easier management of water quality (ODEQ 2013). Refining the scope of the model by scaling the model down to Ecoregion 4 levels, or ODEQS management regions north vs. south, or north, central, and south coast regions may improve accuracy. Final site selection could be explored more, possibly by completely eliminating nested sites, or adapting an approach similarly used in SPARROW nutrient modeling that takes into account an upstream monitoring station being used as an input to downstream sites (Smith et al., 1997). Another possibility for site selection, would to be more stringent on temporal sampling selection, by selecting sites that had heterogeneous seasonality for the study years. Through inclusion of point sources, such as National Pollutant Discharge Elimination System (NPDES) permit sites, known confined feeding operations, applicated sludge locations and quantities, and wastewater treatment plants. Non-point sources such as wildlife could be estimated with tools like Bacteria Source

Load Calculator, or assessing housing residence age or sewerage types along stream ways could add more to direct source inputting (Zeckoski et al. 2005). Integration of the coarser STATSGO (1:250,000) into the missing gaps of the higher resolution SSURGO (1:24,000) could be a viable way to refine soils data. Conversion of the percentage clays, silts, and sands soils types into a more general soil texture as defined by widely used USDA soils triangle could be informative. Using different analytical techniques such as, logistic or generalized linear regression models might also provide improvements over the machine learning used here. The overall process here is sound, and these suggestions and other predictors can be added for another analysis.

Currently, ODEQ is confronted with many TMDL's within the coastal region, and a prediction models like this could be useful in future watershed sampling point selection. Since new data are expensive to obtain, this type of generic approach in analyzing already acquired data could instead be used to inform policy makers and watershed managers of potential problems in Oregon's streams, and provide avenues for predicting future water quality from changing land uses or other anthropomorphic demographics. Models such as this would be useful when fitting TMDL process models, by

highlighting spatial areas and watershed parameters that have the highest influence on bacteria counts. Thus informing model building, fitting, and calibration for mechanistic models during TMDL implementation. The major findings of this research are related to riparian land use, and many partnering organizes are generally focused on riparian restorations efforts in the region. But problems with regional sampling plans remain. Better coordination with stake holder groups that are interested in continued improvements in local water quality means continued improvements in sampling plans, this is where trained scientists at regulatory agencies can help inform the public. Sampling location data tell us that many sites are focused around potential areas of localized concern. But these non-randomized or clustered sampling methods cause problems for researchers and managers trying to apply methodologies to assess a region's water quality. This means difficulty in discovering the syntactical relationships between variables and vectors that protect a stream's water quality. Continued educational outreach to shareholders and the community about water quality problems, research methodologies, and keen awareness of lag times from implementation of best management practices will continue to be key in solving our water quality

issues.

REFERENCES

Abu-Ashour, J., & Lee, H. (2000). Transport of bacteria on sloping soil surfaces by runoff. *Environmental Toxicology,* 15, 149-153.

Barcina, I., Lebaron, P., & VivesRego, J. (1997). Survival of allochthonous bacteria in aquatic systems: A biological approach. *Fems Micriobiology Ecology*, *23*, 1–9.

Becker, W., Nennich, T. D., & Atkinson, S. F. (2010). Survivability of Bovine Derived Escherichia coli Subjected to Temperatures Typical of Summer in Texas. *The Texas Journal of Agriculture and Natural Resource*, *25*, 19–25.

Benham, B. L., C. Baffaut, R. W. Zeckoski, Y. A. Pachepsky, K. R. Mankin, A. M. Sadeghi, K. M. Brannan, M. L. Soupir, & M. J. Habersack. (2006). Modeling bacteria fate and transport in watersheds to support TMDLs. *Transactions of the ASAE*, 49, 987–1002.

Bernhard, A. E., Goyard, T., Simonich, M. T., & Field, K. G. (2003). Application of a rapid method for identifying fecal pollution sources in a multi-use estuary. *Water research*, *37*, 909–13.

Boyer, D., Kuczynska, E., & Fayer, R. (2009) Transport, fate, and infectivity of Cryptosporidium parvum oocysts released from manure and leached through macroporous soil. *Environmental Geology,* 58, 1011–1019

Boyer, D. G., & Neel, J. P. S. (2010). Nitrate and fecal coliform concentration differences at the soil/bedrock interface in Appalachian silvopasture, pasture, and forest. *Agroforestry Systems*, *79*, 89–96.

Chin, D., & Sakura-Lemessy, D. (2009). Watershed-scale fate and transport of bacteria. *Transactions of the ASABE, 52*(, 145–154.

Clarke, S. E., & Schaedel, A. L. (1991). Oregon , USA , Ecological Regions and Subregions for Water Quality Management. *Environmental Management, 15*, 847–856.

Crowther, J, Kay, D., & Wyer, M. D. (2001). Relationships between microbial water quality and environmental conditions in coastal recreational waters: the Fylde coast, UK. *Water research*, *35*, 4029–38.

Crowther, J, Wyer, M. D., Bradford, M., Kay, D., & Francis, C. a. (2003). Modelling faecal indicator concentrations in large rural catchments using land use and topographic data. *Journal of applied microbiology*, *94*, 962–73.

Crowther, John, Hampson, D. I., Bateman, I. J., Kay, D., Posen, P. E., Stapleton, C. M., & Wyer, M. D. (2011). Generic Modelling of Faecal Indicator Organism Concentrations in the UK. *Water*, *3*, 682–701.

Cude, C. G. (2005). Accommodating Change of Bacterial Indicators in Long Term Water Quality Datasets. *Journal of the American Water Resources Association*, 41, 47-54.

Dennehy KF, Litke DW, Tate CM, Qi SL, McMahon PB, Bruce BW, Kimbrough RA, & Heiny JS (1998.) Water quality in the South Platte River basin, Colorado, Nebraska, and Wyoming, 1992–95. *USGS Circular*. 1116.

Desmarais, T. R., Solo-gabriele, H. M., Carol, J., & Palmer, C. J. (2002). Influence of Soil on Fecal Indicator Organisms in a Tidally Influenced Subtropical Environment Influence of Soil on Fecal Indicator Organisms in a Tidally Influenced Subtropical Environment. *Applied and Environmental Microbiology*, 69, 3687-3694.

Doran, J. W., & Linn, D. M. (1979). Bacteriological quality of runoff water from pastureland. *Applied and environmental microbiology*, *37*, 985–91.

D.W. Bauer, D. J. M. & A. C. S. (2002). Streambank slumping and its contribution to the phosphorus and suspended sediment loads of the Blue Earth River, Minnesota. *Journal of Soil and Water Conservation*, 57, 243-250

Ferguson, C., Husman, A. de R., & Altavilla, N. (2003). Fate and transport of surface water pathogens in watersheds. *Critical Reviews in Environmental Science and Technology,* 33, 299-361.

Francis, G. A., & O'Beirne, D. O. (2001). FOOD-BORNE PATHOGENS Effects of vegetable type, package atmosphere and storage temperature on growth and survival of Escherichia coli O157 : H7 and Listeria monocytogenes, *Journal of Industrial Microbiology & Biotechnology*, 27, 111–116.

Frontier, S. (1976). Étude de la décroissance des valeurs propres dans une analyse en composantes principales: comparison avec le modèle du bâton brisé. *J. Exp. Mar. Biol. Ecol*. 25, 67-75.

Garzio-Hadzick, a, Shelton, D. R., Hill, R. L., Pachepsky, Y. a, Guber, a K., & Rowland, R. (2010). Survival of manure-borne *E. coli* in streambed sediment: effects of temperature and sediment properties. *Water research*, *44*, 2753–62.

Guzman, J., Fox, G., Malone, R., & Kanwar, R. (2009). Escherichia coli Transport from Surface-Applied Manure to Subsurface Drains through Artificial Biopores. *Journal of Environmental Quality*, 38, 2412–2421

Gascón, J., Oubiña, A., Pérez-Lezaun, A., & Urmeneta, J. (1995). Sensitivity of selected bacterial species to UV radiation. *Current Microbiology*, *30*, 177–182.

Hevesi, J.A., Flint, L.E., Church, C.D., and Mendez G.O., (2011). Application of a watershed model (HSPF) for evaluating sources and transport of pathogen indicators in the Chino Basin drainage area, San Bernardino County, California: *U.S. Geological Survey Scientific Investigations Report*, 2009-5219, 146.

Jamieson, R. C., Gordon, R. J., Sharples, K. E., Stratton, G. W., & Madani, A. (2002). Movement and persistence of fecal bacteria in agricultural soils and subsurface drainage water : A review. *Canadian Biosystems Engineering,* 44, 1.1-1.9.

Jamieson, R., Gordon, R., & Joy, D. (2004). Assessing microbial pollution of rural surface waters: A review of current watershed scale modeling approaches. *Agricultural water management*, 70, 1-17

Kay, D., Anthony, S., Crowther, J., Chambers, B. J., Nicholson, F., Chadwick, D., Stapleton, C. M., & Wyer (2010). Microbial water pollution: a screening tool for initial catchment-scale assessment and source apportionment. *The Science of the total environment*, *408*, 5649–5656.

Kay, D, Crowther, J., Stapleton, C. M., Wyer, M. D., Fewtrell, L., Anthony, S., Bradford, M., et al. (2008). Faecal indicator organism concentrations and catchment export coefficients in the UK. *Water research*, *42*, 10-11.

Kay, D., Wyer, M., Crowther, J., Stapleton, C., Bradford, M., McDonald, A., Greaves, Fancis, C., & Watkins, J. (2005). Predicting faecal indicator fluxes using digital land use data in the UK's sentinel Water Framework Directive catchment: the Ribble study. *Water research*, *39*, 3967-81.

Kim, J.-W., Pachepsky, Y. a., Shelton, D. R., & Coppock, C. (2010). Effect of streambed bacteria release on *E. coli* concentrations: Monitoring and modeling with the modified SWAT. *Ecological Modelling*, *221*, 1592–1604.

Kim, S. M., B. L. Benham, K. M. Brannan, R. W. Zeckoski, & J. Doherty (2007). Comparison of hydrologic calibration of HSPF using automatic and manual methods. *Water Resources Research*, 43, 1-12.

Kouznetsova, M.Y., Roodsarib, R., Pachepskyc,Y.A., Sheltonc, D.R., Sadeghid, A.M., Shirmohammadib, A., & Starr, J.L (2007). Modeling manure-borne bromide and fecal coliform transport with runoff and infiltration at a hillslope. *Journal of Environmental Management*, 84, 336–346

Kuczynska, E., Boyer, D.G., & Shelton, D.R. (2003). Comparison of immunofluorescence assay and immunomagnetic electrochemiluminescence in detection of Cryptosporidium parvum oocysts in karst water samples. *Journal of microbiological methods*, 53, 17-26.

Leber, J., Rahman, M. M., Ahmed, K. M., Mailloux, B., & van Geen, A. (2011). Contrasting influence of geology on *E. coli* and arsenic in aquifers of Bangladesh. *Ground water*, *49*, 111–23.

Lowrance, R., Altier, L. S., Newbold, J. D., Schnabel, R. R., Groffman, P. M., Denver, J. M., & Todd, A. H. (1997). Water Quality Functions of Riparian Forest Buffers in Chesapeake Bay Watersheds. *Environmental Management*, *21*, 687–712.

Meals, D.W., Dressing, S.A., & Davenport, T.E. (2010). Lag time in water quality response to best management practices: review. *Journal of Environmental Quality,* 39, 85–96.

Mosaddeghi, M., Mahboubi, A., Zandsalimi, S., & Unc, A. (2008). Influence of organic waste type and soil structure on the bacterial filtration rates in unsaturated intact soil columns. *Journal of Environmental Management*, 90, 730-9.

Mubiru, D. N., Coyne, M. S., & Grove, J. H. (2000). Mortality of Escherichia coli O157:H7 in Two Soils with Different Physical and Chemical Properties. *Journal of Environment Quality*, *29*, 1821-1825.

Muirhead, R. (2006). Interaction of Escherichia coli and Soil Particles in Runoff. *Applied and Environmental Microbiology,* 72, 3406-3411.

New Zealand Ministry for the Environment (2010). Modelling water quality in New Zealand rivers from catchment-scale physical, hydrological and land-cover descriptors using random forest models. Retrieved on 4-2-2012 from: http://www.mfe.govt.nz/publications/water/modelling-water-quality-in-nz-rivers/modelling-water-quality-in-nz-rivers.pdf

Oliver, D. M., Heathwaite, A. L., Fish, R. D., Chadwick, D. R., Hodgson, C. J., Winter, M., & Butler, A. J. (2009). Scale appropriate modelling of diffuse microbial pollution from agriculture. *Progress in Physical Geography*, 33, 358-377.

Oregon Department of Environmental Quality (2010) Water Quality Program Rules Retrieved from: http://www.deq.state.or.us/wq/rules/div041tblsfigs.htm

Oregon Department of Environmental Quality (2011) North Coast Water Quality Status & Action Plan: North Coast. Retrieved from: http://www.deq.state.or.us/wq/watershed/Docs/NorthCoastPlan.pdf

Oregon Department of Environmental Quality (2013) Total Maximum Daily Loads (TMDLs) Program. Retrieved from: http://www.deq.state.or.us/wq/t

Osborne, L. L., & Kovacic, D. A. (1993), Riparian vegetated buffer strips in water-quality restoration and stream management. *Freshwater Biology*, 29, 243–258.

Pachepsky, Y. a., Sadeghi, a. M., Bradford, S. a., Shelton, D. R., Guber, a. K., & Dao, T. (2006). Transport and fate of manure-borne pathogens: Modeling perspective. *Agricultural Water Management*, *86*, 81–92.

Paul, M., & Meyer, J. (2001). Streams in the urban landscape. *Annual Review of Ecology and Systematics*, 32, 333-365.

Preston, S.D., Alexander, R.B., Woodside, M.D., and Hamilton, P.A., (2009). SPARROW MODELING—Enhancing Understanding of the Nation's Water Quality: U.S. Geological Survey Fact Sheet 2009–3019, 6 p.

Pronk, M., Goldscheider, N., Zopfi, J., & Zwahlen, F. (2008). Percolation and Particle Transport in the Unsaturated Zone of a Karst Aquifer. *Ground Water,* 47, 361-369.

Rhodes, M. W., & Kator, H. (1988). Survival of Escherichia coli and Salmonella spp. in estuarine environments. *Applied and environmental microbiology*, *54*, 2902–7.

Rozen, Y. and Belkin, S. (2001) Survival of enteric bacteria in seawater. *FEMS Microbiology Review,* 25, 513–529.

Sadeghi, A.M., & Arnold, J. G. (2002) A SWAT/Microbial submodel for predicting pathogen loadings in surface and groundwater at watershed and basin scales (paper 701P0102) Total Maximum Daily Load (TMDL) Environmental Regulations: Proceedings of the March 11-13, 2002 Conference, Fort Worth, Texas, USA. American Society of Agricultural Engineers.

Sekely AC, Mulla DJ, Bauer DW. 2002. Streambank slumping and its contribution to the phosphorus and suspended sediment loads of theBlue Earth River, Minnesota. *Journal of Soil and Water Conservation*, 57, 243–250.

Smith, R. a., Schwarz, G. E., & Alexander, R. B. (1997). Regional interpretation of water-quality monitoring data. *Water Resources Research*, *33*, 2781.

Strobl, C., Hothorn, T., & Zeileis, A. (2009). Party on ! A New, Conditional Variable Importance Measure for Random Forests Available in the party Package Party on ! *The R Journal, 1/2,* 14-17.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, *14*, 323–48

Tate, K. W., Atwill, E. R., Bartolome, J. W., & Nader, G. (2006). Significant Escherichia coli attenuation by vegetative buffers on annual grasslands. *Journal of environmental quality*, *35*, 795–805.

United States Environmental Protection Agency (2012a). National Water Quality Assessment Report. Retrieved from: http://iaspub.epa.gov/waters10/attains_nation_cy.control.

Walters, S. P., & Field, K. G. (2009). Survival and persistence of human and ruminant-specific faecal Bacteroidales in freshwater microcosms. *Environmental microbiology*, *11*, 1410–21.

Wickham, J. D., Stehman, S.V., Gass, L., Dewitz, J., Fry, J., & Wade, T. G. (2013). Accuracy assessment of NLCD 2006 land cover and impervious surface. *Remote Sensing of Environment*, 130, 294 – 304.

Wilkes, Graham, Edge, T., Gannon, V., Jokinen, C., Lyautey, E., Medeiros, D., & Neumann, N. (2009). Seasonal relationships among indicator bacteria, pathogenic bacteria, Cryptosporidium oocysts, Giardia cysts, and hydrological indices for surface waters within an agricultural landscape. *Water research*, *43*, 2209–23.

Wilkes, G, Edge, T. A., Gannon, V. P. J., Jokinen, C., Lyautey, E., Neumann, N. F., & Ruecker, N. (2011). Associations among pathogenic bacteria , parasites , and environmental and land use factors in multiple mixed-use watersheds. *Water Research*, *45*, 5807–5825.

Wilkinson, R., McKergow, L., Davies-Colley, R., Ballantine, D., & Young, R. (2011). Modelling storm-event E . coli pulses from the Motueka and Sherry Rivers in the South Island, New Zealand. *New Zealand Journal of Marine and Freshwater Research*, *45*, 369–393.

Williams, a. P., Quilliam, R. S., Thorn, C. E., Cooper, D., Reynolds, B., & Jones, D. L. (2012). Influence of Land Use and Nutrient Flux on Metabolic Activity of *E. coli* O157 in River Water. *Water, Air, & Soil Pollution*, *223*, 3077–3083.

Wu, J., Rees, P., & Dorner, S. (2011). Variability of *E. coli* density and sources in an urban watershed. *Journal of water and health*, *9*, 94–106.

Zeckoski, R.W., B.L. Benham, S.B. Shah, M.L. Wolfe, K.M. Brannan, M. Al-Smadi, T.A. Dillaha, S. Mostaghimi, and C.D. Heatwole. 2005. BSLC: a tool for bacteria source characterization for watershed management. *Applied Engineering in Agriculture,* 21, 879-889.

SOFTWARE, DATA, AND DATABASES CITED

ESRI, Inc. (2012) ArcGIS 10.0 (Service pack 5). Redlands, California, USA

ESRI, Inc. (2012) USA Population by Zip Code. Redlands, California, USA. Retrieved from: http://www.arcgis.com

Fry, J., Xian, G., Jin, S., Dewitz, J., Homer, C., Yang, L., Barnes, C., Herold, N., and Wickham, J., 2011. Completion of the 2006 National Land Cover Database for the Conterminous United States, PE&RS, Vol. 77(9):858-864. Retrieved: http://www.mrlc.gov/nlcd2006.php

Gesch, D., Evans, G., Mauck, J., Hutchinson, J., Carswell Jr., W.J., 2009, The National Map—Elevation: U.S. Geological Survey Fact Sheet 2009-3053, 4 p. Retrieved from: http://ned.usgs.gov/index.asp

Land Cover Database, New Zealand (2012). Accuracy Assessment, Retrieved from: http://www.lcdb.scinfo.org.nz/about-lcdb/accuracy-assessment

Oregon Department of Environmental Quality (2012) Laboratory Analytical Storage and Retrieval (LASAR) Retrieved from: http://deq12.deq.state.or.us/lasar2/

PRISM Climate Group, Oregon State University. Retrieved from: http://prism.oregonstate.edu

Python Software Foundation (2010). Python (Version 2.6) Beaverton, Oregon, USA. Retrieved from: http://www.python.org/

R Development Core Team. 2012. R: a language and environment for statistical computing. R Foundation for Statistical Computing (Verrsion 2.15.2), Vienna, Austria. Retrieved from: http://cran.r-project.org/

APPENDIX A: GIS MODLES, R AND PYTHON SCRIPTS

This section is intended to detail the geoprocessing and data processing steps taken within the ArcGIS environment, and its built in extension and use of the Python scripting language and R statistics.

```
# Author: Paul Pettus, © 2013 ppettus@pdx.edu ppettus@unzane.com
# R 2.15.2 statistical package
# Purpose: Process zonal statistics for each catchment in the NHDPlus
# Ver 2 dataset.  Land use layer rasters were summed by cell count
# per catchment overlay then saved to .cvs files for each spatial
# layer analyzed

library(raster)
library(rgdal)
library(maptools)
library(foreign)
library(sp)
library(methods)


#RasterLayer with default parameters
# Spatial layers to be processed

nlcd <- raster("C:/WorkSpace/Hydro_Prj/LU_Cl_Catch.tif")
catchments <- raster("C:/WorkSpace/Hydro_Prj/Catch.tif")
nlcd30 <- raster("G:/GIS/Landcover/lc_n83_C_30.tif")
nlcd100 <- raster("G:/GIS/Landcover/lc_n83_C_100.tif")
cattle <- raster("C:/WorkSpace/Hydro_Prj/4_17/cattle.tif")
sheep <- raster("C:/WorkSpace/Hydro_Prj/4_17/sheep.tif")
milk <- raster("C:/WorkSpace/Hydro_Prj/4_17/milk.tif")
chick <- raster("C:/WorkSpace/Hydro_Prj/4_17/chick.tif")
pop <- raster("C:/WorkSpace/Hydro_Prj/4_17/pop.tif")
ele <- raster("C:/WorkSpace/Hydro_Prj/4_17/ele.tif")
slope <- raster("C:/WorkSpace/Hydro_Prj/4_17/slope.tif")
slope30 <- raster("C:/WorkSpace/Hydro_Prj/4_17/slope_30.tif")
slope100 <- raster("C:/WorkSpace/Hydro_Prj/4_17/slope_100.tif")
slope100 <- raster("C:/WorkSpace/Hydro_Prj/4_17/slope_100_2.tif")
slope100 <- raster("G:/GIS/Geology/slope_5-15_degree_clip_100m.tif")
sand <-
raster("G:/GIS/Soils/gsmsoil_or/Attributes_10cm/Raster/Clip/Per_Sand_10
cm_clip.tif")
clay <-
raster("G:/GIS/Soils/gsmsoil_or/Attributes_10cm/Raster/Clip/Per_Clay_10
cm_clip.tif")
silt <-
```

```
raster("G:/GIS/Soils/gsmsoil_or/Attributes_10cm/Raster/Clip/Per_Silt_10
cm_clip.tif")
ksat <-
raster("G:/GIS/Soils/gsmsoil_or/Attributes_10cm/Raster/Clip/Ksat_10cm_c
lip.tif")
awc <-
raster("G:/GIS/Soils/gsmsoil_or/Attributes_10cm/Raster/Clip/AWC_10cm_cl
ip.tif")
sand30 <-
raster("G:/GIS/Soils/gsmsoil_or/Attributes_10cm/Raster/30m/Per_Sand_10c
m_clip_30m.tif")
clay30 <-
raster("G:/GIS/Soils/gsmsoil_or/Attributes_10cm/Raster/30m/Per_Clay_10c
m_clip_30m.tif")
silt30 <-
raster("G:/GIS/Soils/gsmsoil_or/Attributes_10cm/Raster/30m/Per_Silt_10c
m_clip_30m.tif")
ksat30 <-
raster("G:/GIS/Soils/gsmsoil_or/Attributes_10cm/Raster/30m/Ksat_10cm_cl
ip_30m.tif")
awc30 <-
raster("G:/GIS/Soils/gsmsoil_or/Attributes_10cm/Raster/30m/AWC_10cm_cli
p_30m.tif")
sand100 <-
raster("G:/GIS/Soils/gsmsoil_or/Attributes_10cm/Raster/100m/Per_Sand_10
cm_clip_100m.tif")
clay100 <-
raster("G:/GIS/Soils/gsmsoil_or/Attributes_10cm/Raster/100m/Per_Clay_10
cm_clip_100m.tif")
silt100 <-
raster("G:/GIS/Soils/gsmsoil_or/Attributes_10cm/Raster/100m/Per_Silt_10
cm_clip_100m.tif")
ksat100 <-
raster("G:/GIS/Soils/gsmsoil_or/Attributes_10cm/Raster/100m/Ksat_10cm_c
lip_100m.tif")
awc100 <-
raster("G:/GIS/Soils/gsmsoil_or/Attributes_10cm/Raster/100m/AWC_10cm_cl
ip_100m.tif")

precipitation <-
raster("C:/WorkSpace/Hydro_Prj/Climate/ppt_area_catchments.tif")
temp_max <-
raster("C:/WorkSpace/Hydro_Prj/Climate/tmax_area_catchments.tif")
temp_min <-
raster("C:/WorkSpace/Hydro_Prj/Climate/tmin_area_catchments.tif")

processed.LU="C:\\Workspace\\Hydro_Prj\\TabLUArea\\Total\\crosstabLU.cs
v"
processed.LU30="C:\\Workspace\\Hydro_Prj\\TabLUArea\\30Buf\\crosstabLU3
0.csv"
processed.LU100="C:\\Workspace\\Hydro_Prj\\TabLUArea\\100Buf\\crosstabL
U100.csv"

processed.ele="C:\\Workspace\\Hydro_Prj\\TabGeo\\Elevation\\ele.csv"
```

```
processed.slope="C:\\Workspace\\Hydro_Prj\\TabGeo\\Elevation\\slope.csv
"
processed.slope30="C:\\Workspace\\Hydro_Prj\\TabGeo\\Elevation\\slope_3
0.csv"
processed.slope100="C:\\Workspace\\Hydro_Prj\\TabGeo\\Elevation\\slope_
100.csv"
processed.slope100="C:\\Workspace\\Hydro_Prj\\TabGeo\\Elevation\\slope_
100_2.csv"
processed.slope100="C:\\Workspace\\Hydro_Prj\\TabGeo\\Elevation\\slope_
100_4_Gdrive.csv"


# processed .csv files of zonal statistics

processed.pop="C:\\Workspace\\Hydro_Prj\\TabPop\\pop.csv"
processed.sheep="C:\\Workspace\\Hydro_Prj\\TabLivestock\\Sheep\\sheep.c
sv"
processed.milk="C:\\Workspace\\Hydro_Prj\\TabLivestock\\Milk\\milk.csv"
processed.cattle="C:\\Workspace\\Hydro_Prj\\TabLivestock\\Cattle\\cattl
e.csv"
processed.chick="C:\\Workspace\\Hydro_Prj\\TabLivestock\\Chick\\chick.c
sv"

processed.ppt="C:\\Workspace\\Hydro_Prj\\TabClimate\\ppt.csv"
processed.tmax="C:\\Workspace\\Hydro_Prj\\TabClimate\\tmax.csv"
processed.tmin="C:\\Workspace\\Hydro_Prj\\TabClimate\\tmin.csv"

processed.sand="C:\\Workspace\\Hydro_Prj\\TabGeo\\sand.csv"
processed.clay="C:\\Workspace\\Hydro_Prj\\TabGeo\\clay.csv"
processed.silt="C:\\Workspace\\Hydro_Prj\\TabGeo\\silt.csv"
processed.ksat="C:\\Workspace\\Hydro_Prj\\TabGeo\\ksat.csv"
processed.awc="C:\\Workspace\\Hydro_Prj\\TabGeo\\awc.csv"

processed.sand30="C:\\Workspace\\Hydro_Prj\\TabGeo\\sand30.csv"
processed.clay30="C:\\Workspace\\Hydro_Prj\\TabGeo\\clay30.csv"
processed.silt30="C:\\Workspace\\Hydro_Prj\\TabGeo\\silt30.csv"
processed.ksat30="C:\\Workspace\\Hydro_Prj\\TabGeo\\ksat30.csv"
processed.awc30="C:\\Workspace\\Hydro_Prj\\TabGeo\\awc30.csv"

processed.sand100="C:\\Workspace\\Hydro_Prj\\TabGeo\\sand100.csv"
processed.clay100="C:\\Workspace\\Hydro_Prj\\TabGeo\\clay100.csv"
processed.silt100="C:\\Workspace\\Hydro_Prj\\TabGeo\\silt100.csv"
processed.ksat100="C:\\Workspace\\Hydro_Prj\\TabGeo\\ksat100.csv"
processed.awc100="C:\\Workspace\\Hydro_Prj\\TabGeo\\awc100.csv"



#extend raster to both so they match extents
nlcd2<-extend(nlcd, catchments, value=NA)
catchments2<-extend(catchments, nlcd, value=NA)     #extend raster
again

#tabulate crosstab counts of cells (Not Area!)
crossedtabfile <- crosstab(nlcd2, catchments2 , digits=0, long=FALSE,
```

```
useNA="always" )

#flip columns and rows
crossedtabfile2 <- as.matrix(t(crossedtabfile))

#write to csv file
write.csv(crossedtabfile2, file="C:/WorkSpace/crosstabLU.csv")

areacosstab<-function(zones, catch, filelocation){
  rasterextend1<-extend(zones, catch, value=NA)
  rasterextend2<-extend(catch, zones, value=NA)
  crossedtabfile <- crosstab(rasterextend1, rasterextend2 , digits=0,
long=FALSE, useNA="always" )
  crossedtabfile2 <- as.matrix(t(crossedtabfile)) #Flip rows for
columns
  write.csv(crossedtabfile2, file=filelocation)
  return("Done")
  }

areazonalsum<-function(types, catch, filelocation){
  rasterextend1<-extend(types, catch, value=NA)
  rasterextend2<-extend(catch, types, value=NA)
  zonesumtabfile <- zonal(rasterextend1, rasterextend2, fun=sum,
digits=100, na.rm=TRUE)
  #zonesumtabfile2 <- as.matrix(t(zonesumtabfile))
  write.csv(zonesumtabfile, file=filelocation)
  return("Done")
  }
#This is broken see above for fix
areazonalmean<-function(Ltypes, catch, filelocation){
  rasterextend1<-extend(Ltypes, catch, value=NA)
  rasterextend2<-extend(catch, Ltypes, value=NA)
  zonemeantabfile <- zonal(rasterextend1, rasterextend2, fun=mean,
digits=100, na.rm=TRUE)
  #zonesumtabfile2 <- as.matrix(t(zonesumtabfile))
  write.csv(zonemeantabfile, file=filelocation)
  return("Done")
  }


# function allows for processing multiple files at once
# Warning processing large files, and qty's of files is exhaustive
happytimes<-function(){
  #comreturn<-areacosstab(nlcd, catchments, processed.LU)
  #comreturn2<-areacosstab(nlcd30, catchments, processed.LU30)
  #comreturn2<-areacosstab(nlcd100, catchments, processed.LU100)
  #comreturn2<-areazonalsum(pop, catchments, processed.pop)
  #comreturn2<-areazonalsum(chick, catchments, processed.chick)

  #comreturn2<-areazonalsum(sheep, catchments, processed.sheep)
  #comreturn2<-areazonalsum(cattle, catchments, processed.cattle)
  #comreturn2<-areazonalsum(milk, catchments, processed.milk)

  #comreturn2<-areazonalmean(ele, catchments, processed.ele)
```

```
  #comreturn2<-areazonalmean(slope, catchments, processed.slope)
  #comreturn3<-areazonalmean(slope30, catchments, processed.slope30)
  comreturn4<-areazonalsum(slope100, catchments, processed.slope100)
  comreturn4<-areazonalmean(slope100, catchments, processed.slope100)

  #comreturn4<-areazonalmean(precipitation, catchments, processed.ppt)
  #comreturn4<-areazonalmean(temp_max, catchments, processed.tmax)
  #comreturn4<-areazonalmean(temp_min, catchments, processed.tmin)

  #comreturn4<-areazonalmean(sand, catchments, processed.sand)
  #comreturn4<-areazonalmean(sand30, catchments, processed.sand30)
  #comreturn4<-areazonalmean(sand100, catchments, processed.sand100)

  #comreturn4<-areazonalmean(clay, catchments, processed.clay)
  #comreturn4<-areazonalmean(clay30, catchments, processed.clay30)
  comreturn4<-areazonalmean(clay100, catchments, processed.clay100)

  #comreturn4<-areazonalmean(silt, catchments, processed.silt)
  #comreturn4<-areazonalmean(silt30, catchments, processed.silt30)
  comreturn4<-areazonalmean(silt100, catchments, processed.silt100)

  #comreturn4<-areazonalmean(ksat, catchments, processed.ksat)
  #comreturn4<-areazonalmean(ksat30, catchments, processed.ksat30)
  comreturn4<-areazonalmean(ksat100, catchments, processed.ksat100)

  #comreturn4<-areazonalmean(awc, catchments, processed.awc)
  #comreturn4<-areazonalmean(awc30, catchments, processed.awc30)
  comreturn4<-areazonalmean(awc100, catchments, processed.awc100)

  }

h2<-happytimes()
```

```
# Author: Paul Pettus, © 2013 ppettus@pdx.edu ppettus@unzane.com
# Python 2.6
# Purpose: Generate a to:from catchment data dictionary list for each
# catchment in NHDPlus PlusFlow.dbf database.  This dictionary list
# can then be used to aggregate catchment attributes

# Import system modules
import os, csv
from collections import deque, defaultdict

def children(token, tree):
#    "returns a list of every child"
    #print ("Token:", token)
    visited = set()
    to_crawl = list([token])
    to_crawl2 = list([])
    #to_crawl = deque([token])
    #print (visited)
```

```
    while to_crawl:
        current = to_crawl.pop()  #was .popleft
        if current in visited:
            continue
        to_crawl2.append(current)
        visited.add(current)
        node_children = set(tree[current])
        to_crawl.extend(node_children - visited)   #was .extendleft
        #to_crawl2.append(node_children - visited)
        #print("visited:",visited)
    #testdic = dict()
    #testdic = visited
    #return (testdic)
    #print ("to_crawl2:",to_crawl2)
    #print ("list(visited): ",list(visited))
    #return list(visited)
    return (to_crawl2)

Flow = dict()
#walking the NHDPlus Flow table
#rows =
arcpy.SearchCursor("G:/GIS/NHDPlus/NHDPlusPN/NHDPlusPN/NHDPlus17/NHDPlu
sAttributes/PlusFlow.dbf")

PlusFlow = ("C:/WorkSpace/Hydro_Prj/4_17/PlusFlow.csv")
#PlusFlow = ("C:\\GIS\\Workspace\\PF.txt.txt") #epa
#PlusFlow = ("C:/WorkSpace/Hydro_Prj/4_17/test_to_from.csv")
#rows = open(PlusFlow, 'r')


with open(PlusFlow, 'rb') as csvfile:
    spamreader = csv.DictReader(csvfile, delimiter=',')
    for row in spamreader:
        ToCOM = row['TOCOMID']
        #print(row['TOCOMID'])
        FromCOM = row['FROMCOMID']
        if int(ToCOM) and int(FromCOM) != 0:
            Flow[FromCOM] = ToCOM
    #print("Done finding")
        #print("Found Line")

d2 = defaultdict(list)
for k,v in Flow.items():
    d2[v].append(k)
Full_Flow = dict()
for items in d2.keys():
    Full_Flow[items] = children(items, d2)
#print(Full_Flow)
print("Done Full_Full")

#outfile = open("C:/WorkSpace/Hydro_Prj/4_17/test.csv",'w')
#outfile2 = open("C:/GIS/WorkSpace/walk_test2.csv",'w') #epa
outfile2 = open("C:/WorkSpace/Hydro_Prj/4_17/walk_test_5-24.csv",'w')
#home
```

```
print("outfile opened")

#infile= ("C:/GIS/WorkSpace/catchments.csv", 'r') #EPA
infile= ("C:/WorkSpace/Hydro_Prj/4_17/catchments.csv") #home
#infile= ("C:/WorkSpace/Hydro_Prj/4_17/catch_test.csv")
print("infile opened")

catchments = [str(line.rstrip()) for line in open(infile, 'r')]
print("Input catchments")

#infile = csv.reader("C:/GIS/WorkSpace/catchments.csv", delimiter=',')
#outfile.write("HUC\n")


#data = ["value %d" % i for i in range(1,4)]

out = csv.writer(outfile2, delimiter=',', lineterminator='\n')


#for i in catchments:
#    outfile.write(str(i))
#    outfile.write(",")
#    outfile.write("\n")
print("Starting catchments")

for i in catchments:
    x = list(i)
    #x.append('\n')
    #print(x)
    value = Full_Flow.get(i)
    #print(value)
    if str(value) == 'None':
        print (i)
        print("We found a None")
        #outfile.write(str(x))
        #outfile.write(str(x))
        #outfile.write(",")
        #outfile.write("\n")
        #print("But added it any ways")
        #data = ["value %d" % i for i in range(x)]
        #out.
        out.writerow([i])
    else:
        hucs = Full_Flow[i]
        #print(hucs)
        type(hucs)
        #for huc in hucs:
            #outfile.write(str(huc))
            #outfile.write(",")
        out.writerow(hucs)
        #outfile.write("\n")
        #print("Successful run through HUCS. We added i ...")
outfile2.close()
print("Done")
```

```
#catchments.close()
print("Done")


# Author: Paul Pettus, © 2013 ppettus@pdx.edu ppettus@unzane.com
# Python 2.6
# Purpose: From the to:from catchment data dictionary, created
# in the previous script this dictionary list
# can then be used to aggregate catchment attributes.
# Each catchment is weighted by it total contributing area
import sys, os, csv

#LU30 = open("C:/WorkSpace/Hydro_Prj/4_17/test.csv",'r')
#LU100 = open("C:/WorkSpace/Hydro_Prj/4_17/test.csv",'r')

#*****************************************************************
**************************************************#
#This fills the watershed characteristic dictionary
#LUinfile = open("C:/WorkSpace/Hydro_Prj/5_1/crosstabLU.csv",'r')
#Probably be best to make one dictionary with all attributes
LUinfile = open("E:/Python/Input/catch_test_LU.csv",'r')
Landuse = csv.DictReader(LUinfile)

LU_Sums = {}
for row in Landuse:
    key = row.pop('Catch_ID')
    if key in LU_Sums:
        # implement your duplicate row handling here
        pass
    LU_Sums[key] = row
#print LU_Sums
#test = list(result.keys())
#*****************************************************************
**************************************************#


#*****************************************************************
**************************************************#
#For each row of the collective catchment file
#look up each catchment in the land use stats file
#sum the area
#create area weight based on total catchments

FLOWinfile = open("E:/Python/Input/test2.csv",'r')
#FLOWinfile = open("C:/WorkSpace/Hydro_Prj/5_1/test_5_1.csv",'r')
#FLOWinfile.next()  #Needed to move past first line

#outfile = open("C:/WorkSpace/Hydro_Prj/4_17/WeightedCatches.csv", 'w')
outfile = open("E:/Python/Input/WeightedCatches2z.csv", 'w')
outfile.write("Catch_ID,")
#fileHeaderlist= list(LU_Sums['23735707'])  # Creating column headers
########## CHange this back too!!!!!
fileHeaderlist= list(LU_Sums['1'])  # Creating column headers
########## CHange this back too!!!!
```

```
for eachcolumn in fileHeaderlist:
    outfile.write(eachcolumn)
    outfile.write(',')
outfile.write('\n')

headlength = []
for count in fileHeaderlist:
    headlength.append(0)

for line in FLOWinfile:
    parts = line.split(',')
    catchmentNumbers = [int(L) for L in parts]
    HUC_ID = 0
    allCells = 0
    matrix = [fileHeaderlist,headlength]
    p1 = 0
    for catchment in catchmentNumbers:
        print ("catchment loop", catchment)
        value = LU_Sums.get(str(catchment))
        if allCells == 0:
            HUC_ID = catchment
        catch_stats = LU_Sums[str(catchment)]
        cells = 0
        p2 = 0
        for (k,v) in catch_stats.items():
            matrix[1][p2] = matrix[1][p2] + int(v)
            cells = cells + int(v)
            p2 = p2 + 1
        allCells = allCells + cells
    outfile.write(str(HUC_ID))
    for x in fileHeaderlist:
        outfile.write(',')
        outfile.write(str(matrix[1][p1]))
        p1 = p1 + 1
    outfile.write('\n')
outfile.close()
#*****************************************************************
*****************************************************#

print ("DONE")
```