ELSEVIER

Scopus

doi

Journal of Vibration Engineering

ISSN:1004-4523

Registered

- SCOPUS
- GOOGLE SCHOLAR
- DIGITAL OBJECT IDENTIFIER (DOI)
- IMPACT FACTOR 6.1

Our Website
www.jove.science

Designing a framework for implementing CUSUM algorithms for Climatechangedetectionforrelativediseases

Ms.AbhiruchiMishra[1]
[1]Student,DepartmentofComputerScienceandEngineering,
[1]ShriSantGajananMaharajCollegeOfEngineering,

Abstract— Big data have long predicted widespread irresistible infections as large-scale reactions to seasonal climate change,polarisingdiscussion,anda particularly dreadfulhumanvirus forwhichfinancialmotivesandcontroltechniquesmightrestrict the disclosure of atmospheric intervention changes. The majority of the time, seasonal climate change illnesses are nolongercommonbecauseofmodificationstorurallandpractises,isolatedincidents,changesinhumanconduct,andmanagementofvectors.Mostvector-borneillnessesonlyseldomshowsymptoms,andtheyareclimatesensitive.TheCumulative Sum (CUSUM) algorithms are used in Big Data Analytics to identify climate change. The recommended CUSUMalgorithmproducesbetterresults for the climate changealgorithm.

Keyterms--Climatechange,Bigdata,Cusum,Dataanalytics.

IIntroduction

To predict climatic changes and disease from the environment, large-scale climate data have been used. It integratesseveral data processing prototypes, including graph-based technique, closest neighbour algorithm, principal componentanalysis as an unsupervised model, binary segmentation models, and lastly a spatial autocorrelation model, to boostscalability and feasibility. The aforementioned strategies are thoroughly analysed in the proposed system to offer a novelparadigm for handling complex climatic data. The analysis of the solution to managing complex data has taken intoaccount a number of aspects, such as creating a trajectory for data flow, projecting the load balancing model for datamanagement, cooperating to handle conjunctions, and cooperating to anticipate uncertainty. It is very difficult to predictclimate changeand its potential effects. [1]

Numerous studies have been conducted on big data analytics-based climate modelling. Data on the climate and seasonalchanges are gathered, shared, and monitored by the US Environmental Security Agency. These have been applied toevaluate the results of ongoing healthcare reforms. Complex big data solutions must be used to process enormous datasources in the healthcare sector. Recent state-of-the-art study concentrates on how big data impacts disease diagnosis. Thesignificance of the massive data stored in big data and cloud computing has also increased noticeably.Advanced big datasolutionswillhelpinthediagnosisandprognosisofdiseases,aswellasinthereductionoftreatmentcostsandimprovement of healthcare quality. One of the algorithms used for one of the approaches to predict climate change is theCUSUM Algorithm. Big data analytics is becoming more and more crucial in a variety of industries, including healthcare,social networking, and weather forecasting. The state-of-the-art sensor can be used to collect climate data, which depictsseasonal changes. Weather information can be used to research seasonal diseases and follow seasonal fluctuations, whilemeteorologicaldataareutilisedto forecasttheweather.

Analysing whether or not the average daily climate compares favourably to the previous climatic time period is a crucialstep in comprehending seasonal changes. Weather information acquired from many sources has been used to identifyseasonalclimaticvariations.Unstructureddatapatternsweremostlyusedtoevolveclimatetesting.Thekeycomplementary situations that raise mathematical problems in a multidisciplinary context include epidemiology, health,biodiversity, natural resources, and seasonal climate change. In order to execute complex workflow pipelines and astrategyformanagingheterogeneousandhugedatasets,somestatedscenariosdemandaccesstotheprovidedframework.[2]

Brownian motion drift changes are the main subject of the first exact optimality result in sequential change detection.UsingaBayesianapproach,thechangetimeistreatedasanindependentrandomvariablewithanexponentialdistribution. The alternative detection approach exhibits significant similarities to the greatest Bayesian test yet developed,while being developed using a minimax strategy. Unfortunately, it has been demonstrated that the Shiryaev-Roberts-Pollak (SRP) test, a sequential detector, is wrong while having a very strong asymptotic optimality property. Also worthnoting is how easy and quick it is to compute the associated test statistic for the CUSUM and SRP tests in the classicalversion oftheproblemthatwehave sofarlooked at.

1.1 Objective

- ToimplementtheCUSUMalgorithm andLogisticRegression.
- Toanalyzetheinputdata forHandling Missing data,LabelEncoding anddropping unnecessary columns

1.2 Problemstatement

Online change point detection is a fundamental problem in statistics and signal processing which finds applications in aplethora of practical problems in diverse fields. The most common version of the problem consists of a sequence ofobservations sampled independently. There is also a change point such that the underlying distribution changes from onedistributiontoanalternative.Thisproblemisofmajorimportanceinmanyapplications,suchasseismicsignalprocessing, industrial quality control, dynamical systems monitoring, structural health control, event detection, anomalydetection, detection of attacks, etc. The goal of online change point detection is to detect the occurrence of the change instatistical behavior with a minimal delay while controlling the false alarm rate. The suitable tradeoff between detectiondelay and false alarm rate, as in all detection problems, is of essential importance for the proper mathematical formulationof the problem. Online change detection involves monitoring a stream of data for changes in the statistical properties ofincoming observations. A good change detector will detect any changes shortly after they occur, while raising few falsealarms.Althoughthereareal gorithmswithconfirmedoptimalitypropertiesforthistask,theyrelyontheexactspecificationsoft herelevant probabilitydistributionsandthislimitstheirpracticality.

IILiteratureSurvey

Nandhini V et. al. states that Big data analytics has been applied to predict climate changes and climate borne diseasefromlargescale climate data.It incorporates various data processingprototypesuch as linearand non-linear dataprocessing algorithm, binary segmentation models, nearest neighborhood algorithm, principal component analysis as anunsupervised model, Graph-based technique and finally spatial autocorrelation model to increase the scalability andfeasibility. In this paper, a detailed study is carried out on the above-said methods to devise a novel paradigm to handlecomplex climate data. The solution to managing complex data has been analyzed on a various aspect such as constructingtrajectory for data flowing, projecting the load balancing model for data managing, collaborative model for conjunctionshandling and joint model for uncertainty prediction. Big data analytics is improving a lot of importance in healthcare,social networking, climate modeling and so on. Climate data reflect seasonal changes, and it could be collected using theadvanced sensor. Weather prediction is made using metrological data, and weather data also useful to analyze seasonaldiseases, and reflects seasonal changes. This helps in improving public health. Usually, the climate data is raw, and it isunstructured format from this data meaningful information extracted using analytical techniques. Recently, roboticsalgorithms are also found useful for climate data analyses. The collection of a vast amount of weather data from overallclimate system. Climate "normal" are used to compare the climate condition of present and past. The average climateparameter ("maximumorminimum temperature") during a periodis usedtocalculatenormalclimate.[1]

LiyanXieet. al. developed an alternative approach for solving the problem of interest which we call Window-LimitedCUSUM (WLCUSUM). It consists in adopting a window-based estimate of the unknown post-change parameters and,unlike the existing window-limited GLR, we use the estimate in the updating formula of the classical CUSUM statistic.This updating mechanism is far more efficient than its window-limited GLR counterpart and by proper selection of thewindowsizewecanalsoguaranteeasymptoticoptimality.Wewouldliketoemphasizethattheproblemweconsiderisnot jointdetectionandestimation, where thetwo tasks areregardedas equally important. Here, we are primarilyinterested in detection, with estimation being an auxiliary action that contributes towards our detection goal. For thisreason we only require the estimator to be consistent without insisting on any explicit form. Compared with existingCUSUM-like procedures employing estimates of post-change parameters, the proposed WLCUSUM method applies to afar wider range of parametric distributions and not only the exponential family which is mostly the case with the availableapproaches.

Thomas et. al.introduced the Kernel Cumulative Sum Algorithm (KCUSUM), a new approach for online changedetection. Unlike the CUSUM algorithm, this approach does not require knowledge of the probability density ratio for itsimplementation. Instead, it uses incomingobservations and samples fromthe pre-change distribution.The result is thatthe same algorithm works for detecting many types of changes. Our theoretical analysis establishes the algorithm's abilityto detect changes, and shows a relation between the delay and the MMD distance of the two distributions. These boundsshould also be useful in the analysis of other non-parametric change detectors. Finally, we would like to suggest twoavenues for future work. First, there are likely variants of KCUSUM that leverage more complex that may lead toimproved detection performance. Secondly, the CUSUM has been investigated for detecting changes in scenarios withmore complex dependencies among observations and it may also be possible to extend the kernel methods developed inthis papertodetectchangesinthese cases.

In this study the change detection problem in a general HMM when the change parameters are unknown and the changecan be slow or drastic. Drastic changes can be detected easily using the increase in tracking error or the negative log ofobservation likelihood (OL). But slow changes usually get missed. We have proposed in past work a statistic called ELLwhich works for slow change detection. Now single time estimates of any statistic can be noisy. Hence we propose amodification of the Cumulative Sum (CUSUM) algorithm which can be applied to ELL and OL and thus improves bothslow and drastic change detection performance. Change detection is required in many practical problems arising in qualitycontrol,flightcontrol,andfaultdetectionandinsurveillance problemslikeabnormalactivitydetection.Inmost cases,the

underlying system in its normal state can be modelled as a parametric stochastic model which is usually nonlinear. Theobservations are usually noisy (making the system partially observed). Such a system forms a "general HMM" (alsoreferred to as a "partially observed nonlinear dynamical model" or a "stochastic state space model" in different contexts).The General Hidden Markov Model library (GHMM) is a freely available C library implementing efficient data structuresand algorithms for basic and extended HMMs with discrete and continuous emissions. It comes with Python wrapperswhich providea muchnicerinterface andaddedfunctionality.

Namrataet. al. Given its associated burden of disease, climate change in South Africa could be reframed aspredominately a health issue, one necessitating an urgent health-sector response. The growing impact of climate changehas major implications for South Africa, especially for the numerous vulnerable groups in the country. We systematicallyreviewed the literature by searching PubMed and Web of Science. Of the 820 papers screened, 34 were identified thatassessed the impacts of climate change on health in the country. Most papers covered effects of heat on health or oninfectious diseases (20/34; 59%). We found that extreme weather events are the most noticeable effects to date, especiallydroughts in the Western Cape, but rises in vector-borne diseases are gaining prominence. Climate aberration is also linkedinmyriadwayswithoutbreaksof foodand waterbornediseases, and possibly withtherecentListeriaepidemic.

IIIProposedSystem(MethodologyandTechniques)

PythonPackages:

- **ocpdet**

OCPDet is a Python package for online changepoint detection, implementing state-of-the-art algorithms and a novelapproach.

- **Itertools**

This module implements a number of iterator building blocks inspired by constructs from APL, Haskell, and SML.Each has been recast in a form suitable for Python. The module standardizes a core set of fast, memory efficient toolsthat are useful by themselves or in combination. Together, they form an "iterator algebra" making it possible toconstructspecializedtoolssuccinctlyand efficiently inpurePython.

- **Numpy**

NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra,fouriertransform,andmatrices.UsingNumPy,mathematicalandlogicaloperationsonarrayscanbeperformed.This tutorial explains thebasics ofNumPysuchas its architectureandenvironment. Italso discusses thevariousarray functions, types of indexing, etc. An introduction to Matplotlib is also provided. All this is explained with thehelpofexamples forbetterunderstanding.

- **Pandas**

Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, andmanipulating data. Pandas allow us to analyze big data and make conclusions based on statistical theories. Pandas canclean messy data sets,andmakethem readableandrelevant.

Methodology

- **Inputdata:**Here,wecancollecttheclimatechangesouthAfricaDatasetfromdatasetrepository.Thedatasetisinthe format'.csv'or'.xlxs'.
- **Pre-processing**:Here,the collectedinputdata iscarried outtopre-processing step,
- HandlingMissingdata
- LabelEncoding
- Dropunnecessarycolumns
- **DataSplitting**:In thisstep,wecan splitthepre-processeddatainto testand trainfordecisionmaking,
- Testdataisusedforprediction (30%)
- Train data is usedfortraining(70%)
- **Classification**:Here, wecan implementthe twodifferentregression algorithmssuch as
- CusumAlgorithm
- Linear Regression (LR)
- **Forecasting:**Here,we canforecastoranalyse the**climate change**(temperature)based on dataset attributesbyusing regressionalgorithms.
- **PerformanceEstimation**:Here,wecanestimatesomeperformancemetricssuchas
- MeanAbsoluteError
- MeanSquaredError
- RootMeanSquaredError
- ForecastingGraph

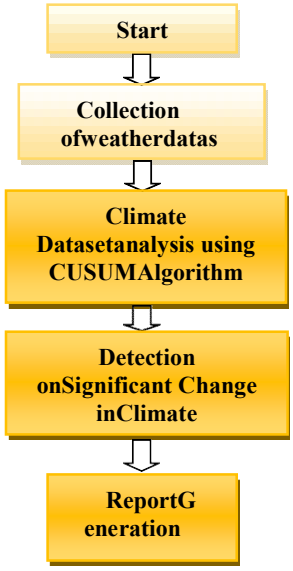

Fig.1 FlowChartforDetectionofClimateChange

CUSUMAlgorithm

A cumulative sum chart (CUSUM), a type of control chart, is used to detect the divergence of the individual values orsubgroup mean from the changed target value or to monitor the variation from the target value. CUSUM charts can beused in place of Shewhart control charts. The CUSUM chart has the advantage of being more sensitive to the little shift ofthe process mean when compared to the Shewhart charts (Individuals I-MR or Xbar charts).The cumulative sum chart andthe exponentially weighted moving average (EWMA) chart both display the process' mean, but they differ significantlyfrom Xbar charts in that they also display the means of the previous values at each point. These charts are regarded asaccurate approximations when the correct standard deviation is included. Instead of calculating the subgroup meansindependently, the data from the current and previous samples is displayed using a CUSUM chart. As a result, theCUSUMchart isconsistentlybetter thanthe Xbarplotsat identifying minute changesin processmean.

LogisticRegression

Logistic regression determines the probability that an event, such as voting or not voting, will occur based on a collectionof independent variables. Since the outcome is a probability, the range of the dependent variable is 0 to 1. The odds, orlikelihoodof successdividedby probabilityoffailure, aretransformedusing thelogitformulain logisticregression. Inthecontext ofartificialintelligence,thesupervisedmachinelearningmodelfamilyincludeslogisticregression.Additionally, it is recognised as a discriminatory model, which means that it tries to differentiate across classes. It cannot,contrary to a generative algorithm like naive bayes, generate information of the class that it is trying to predict (forexample, a picture of a cat). Essentially, logistic regression does supervised categorisation. In a classification problem, thegoal variable (or output), y, can only take discrete values for a particular set of features (or inputs), X. Contrary to popularbelief, a logistic regression is a regression model. The model builds a regression model to predict the probability that aparticular data entry will fall into the group denoted by the number "1". Similar to how linear regression presumes that thedata follows a linear distribution, logistic regression models the data using the sigmoid function. Logistic regression onlybecomes a classification strategy when a decision threshold is added. The logistic regression's key element, the thresholdvalue, is determined by the classification problem itself. The threshold value selection is significantly influenced by theprecision and recall levels. Precision and recall ought to be equal in a perfect environment, but this is rarely the case.Thesetup of the reference levels must therefore be carefully considered. If there is no obvious rule generated from the dataitself or by prior knowledge about the variable values, it is advised to set a reference level with a minimum sample size toprovide appropriate statistical power.Another approach to aid understanding is to select categories that have a similarconnection to the pertinent event.If you believe that those who are older and those who are undergoing innovativetreatments havelower mortalityrates,usethesetwogroups as apoint ofcomparison.Alternately, you might useconventional medical care and younger patients. It is doable and not inaccurate to choose older individuals and receiveregular care,butdoing sowillmake itmorechallengingtocomprehend theresults.

IVPerformanceAnalysis

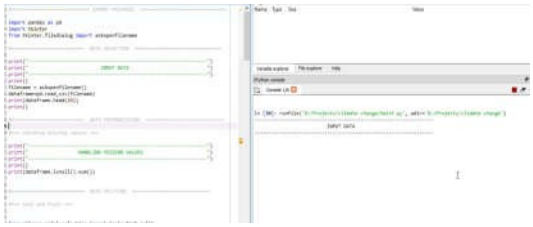

Fig.1Editor Window

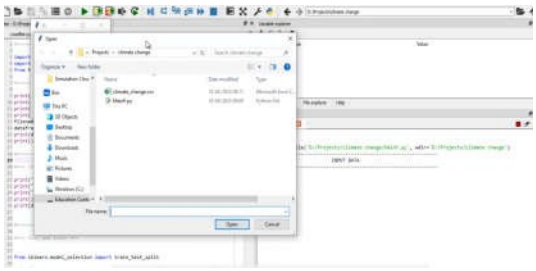

Fig2Selectionofdataset

```
----------------------------------------------------------------
                    DATA SPLITTING
----------------------------------------------------------------

Total No. of data's in input              : 308

Total No. of data's  in training data     : 246

Total No. of  data's in testing data      : 62

----------------------------------------------------------------
                    LINEAR REGRESSION
----------------------------------------------------------------

1. Mean Absolute Error ( MAE)     : 0.07393626391432558

2. Mean Squared Error (MSE)       : 0.008097773547311104

3. Root Mean Squared Error (RMSE)  : 0.08998762996829678
```

Fig3Datasplittingandlinearregression

```
LONG SHORT TERM MEMORY ---> LSTM

Train on 277 samples, validate on 31 samples
Epoch 1/5
277/277 [==============================] - 9s 33ms/step - loss: 0.1208 - mae: 0.2571 - val_loss:
0.0890 - val_mae: 0.2821
Epoch 2/5
277/277 [==============================] - 1s 2ms/step - loss: 0.0514 - mae: 0.1811 - val_loss:
0.0106 - val_mae: 0.0812
Epoch 3/5
277/277 [==============================] - 1s 2ms/step - loss: 0.0550 - mae: 0.1882 - val_loss:
```

Fig4 ApplicationofLSTM

```
                    PREDICTION
----------------------------------------------------------------

================================================

[0] The Climate Change= -0.007342379340855132
================================================

[1] The Climate Change= 0.42810863032528346
================================================

[2] The Climate Change= 0.3143244543422554
================================================

[3] The Climate Change= 0.18759464295149542
================================================

[4] The Climate Change= 0.33099758770421627
```

Fig5Climatechange prediction

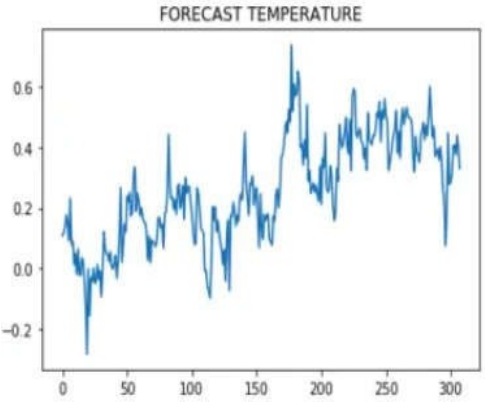

Fig6Climateforecasting fortemperature

VConclusion

From the perspective of practical and precise calculation, CUSUM is a costly step for detecting climate changes.Thesuggestedapproachlooksforchanges brought onbyclimatechangeusingthe CUSUMalgorithm.Utilisingaclimatic data collection, the performance of the recommended solution is confirmed. A data cleaning technique is used todecrease the amount of data that needs to be collected, increase the accuracy of the current changes in climatic conditions,andsavetimeand money. Themoderateandradicalchanges inthemeanassessment ofaquantityof intrigue areidentified using seasonal fluctuations and aggregate approaches. This method is used in a variety of activities; it shows onthescreenchangesingenerationalconditions,diseasedesires,fishnumbers,deforestation,andinvestigationsintowrongdoing.
Finally, we would like to suggest two potential lines of inquiry for further investigation. First, it's likely thatthere are KCUSUM variants that use more complex non-i.i.d. statistics and could improve detection performance. Second,theCUSUMhasbeenstudied in contextswithmorecomplex relationshipsbetweenobservationsforchangedetection.

REFERENCES

[1] V.Nandhini, Dr.M.S.GeethaDevasena, "Predictive Analytics for Climate Change Detection and Disease Diagnosis", 20195th InternationalConferenceonAdvanced Computing&CommunicationSystems(ICACCS)

[2] LiyanXie, George V. MoustakidesAnd YaoXie ," Window-Limited CUSUM for SequentialChange Detection",arXiv:2206.06777v1[math.ST]14Jun2022

[3] Thomas Flynn and ShinjaeYoo," Change Detection with the Kernel Cumulative Sum Algorithm", 2019 IEEE 58thConferenceonDecisionandControl(CDC)PalaisdesCongrèsetdesExpositionsNiceAcropolisNice,France,December 11-13,2019.

[4] NamrataVaswani," The Modified Cusum Algorithm For Slowand Drastic Change Detection In General Hmms WithUnknownChange Parameters",0-7803-8874 7/05/©2005IEEE ICASSP2005

[5] Matthew F. ChersichCaradee Y. Wright, Francois Venter, Helen Rees, Fiona Scorgie, and Barend Erasmus," Impacts ofClimateChangeonHealthandWellbeinginSouthAfrica,Int.J.Environ.Res.PublicHealth2018,15,1884;doi:10.3390/ijerph15091884www.mdpi.com/journal/ijerph

[6] C.Yu,R.Zhang,Y.Huang,andH.Xiong,"High-dimensionalknnjoinswithincrementalupdates,"GeoInformatica,2010.

[7] W.Welcome,"http://www.wmo.int/pages/prog/wcp/wcdmp/index_en.php,"2017.

[8] LeeJG,KangM,"Geospatialbigdata:challengesandopportunities,"BigDataRes,pp.74–81,2015.

[9] NativeS,MazzettiP,SantoroM,PapeschiF,CragliaM,OchiaiO,"Bigdatachallengesinbuildingtheglobalearthobservationsystemofsystems,"Environ ModelSoftw, vol.68,pp.1–26,2015.

[10] EnviroAtlas|USEnvironmentalProtectionAgency,http://enviroatlas.epa.gov/enviroatlas,2017.

[11] Pickard BR, Baynes J, Mehaffey M, Neale AC, "Translating big data into big climate ideas, "Solutions, vol.6,pp.64–73,2015.

[12] Lopez D, Gunasekaran M, Murugan BS, Kaur H, Abbas KM, "Spatial big data analytics of influenza epidemic in Vellore,IEEEinternationalconferenceonbigdata(bigdata),"IEEE, vol.27, pp.19–24,2014.

[13] LopezD,GunasekaranM,"Assessmentofvaccinationstrategiesusingfuzzymulticriteriadecisionmaking.In:Proceedings of the fifth international conference on fuzzy and neurocomputing (FANCCO-2015)," Springer InternationalPublishing,pp.195–208,2015.

[14] LopezD,SekaranG,"Climatechangeanddiseasedynamics-abigdataperspective.IntJInfectDis,"pp.23–4,2016.

[15] Lopez D,ManogaranG,"Bigdata architecturefor climate changeanddiseasedynamics, "The human elementofbig data :issues,analytics,and performance.USA:CRCPress;2016.

[16] Manogaran G, Thota C, Kumar MV, " MetaCloudDataStorage architecture for big data security in cloud computing,"ProcediaComputSci,pp.128–33,2016.

[17] ManogaranG,ThotaC,LopezD,VijayakumarV,AbbasKM,SundarsekarR,"Bigdataknowledgesysteminhealthcare.In:Internet ofthingsandbigdatatechnologiesfornext-generationhealthcare,"SpringerInternationalPublishing,pp.133–57,2016.

[18] Manogaran G, Lopez D. Disease surveillance system for big climate data processing and dengue transmission. Int JAmbientComputIntell,pp.88–105,2017.

[19] Thota C, Manogaran G, Lopez D, Vijayakumar V, "Big data security framework for distributed cloud data centers.In:Cybersecuritybreachesandissuessurrounding onlinethreatprotection,"USA:IGIGlobal,pp.288–310, 2017.

[20] Manogaran G, Thota C, Lopez D, Vijayakumar V, Abbas KM, Sundarsekar R, "Big data knowledge system in healthcare.In: Internet of things and big data technologies for next-generation healthcare," Springer International Publishing, pp.133–57,2017.

[21] MurdochTB,Detsky AS,"Theinevitableapplicationofbigdata tohealthcare,"vol.309(13,pp.1351–2,2013.

[22] Haines, A.; Kovats, R.S.; Campbell-Lendrum, D.; Corvalan, C. Climate change and human health: Impacts, vulnerabilityand publichealth.PublicHealth 2006,120,585–596.[CrossRef][PubMed]

[23] Cohen, A.J.; Brauer, M.; Burnett, R.; Anderson, H.R.; Frostad, J.; Estep, K.; Balakrishnan, K.; Brunekreef, B.; Dandona,L.; Dandona, R.; et al. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution:Ananalysisof data fromtheGlobalBurdenofDiseasesStudy 2015.Lancet2017, 389,1907–1918. [CrossRef]

[24] MacKellar,N.;New,M.;Jack,C.ClimatetrendsinSouthAfricaObservedandmodelledtrendsinrainfallandtemperature forSouth Africa:1960–2010.S.Afr.J.Sci.2014,110,1–13.[CrossRef]

[25] Van Wilgen, N.J.; Goodall, V.; Holness, S.; Chown, S.L.; Mcgeoch, M.A. Rising temperatures and changing rainfallpatternsin South Africa'snationalparks.Int.J.Climatol.2016,36,706–721.