



# Draftable Comparison Export

This document is an exported comparison with limited functionality, generated by Draftable Desktop. To access full functionality, use Draftable's powerful comparison viewer in any of our products.

**Left document:** biody\_informatics.pdf

**Right document:** biody\_informatics\_rev1.pdf

## What is this document?

This is a comparison of two documents. The two documents are interleaved such that the left document is displayed on even pages and the right document is displayed on odd pages.

## Is there a specific way I should view this file?

This document is intended to be viewed in Two Page Continuous mode (or sometimes called 'Two Page Scrolling'). It should open in this mode by default when using Adobe Acrobat and most popular PDF readers.

If the document opens in a different view, you can often change this in the settings. In Adobe Acrobat, go to **View > Page Display > Two Page Scrolling**.

## Why are there blank pages?

Blank pages are inserted to keep both documents as aligned as possible.

## How do I read the changes?

Text deleted from the left document and, hence, not in right document is highlighted red. Text added to the right document and, hence, not in left document is highlighted green.

## Tip for printing

When printing this document, we recommend printing double-sided and include this first page. This will result in the matching text being displayed on different pages and easily readable, much like a book.

## For more information

Draftable offers powerful document comparison solutions for all use-cases. To view our products, please visit our website: [draftable.com](http://draftable.com).

# BIRDIE: A data pipeline to inform wetland and waterbird conservation at multiple scales

Francisco Cervantes<sup>1,2\*</sup>, Res Altwegg<sup>1</sup>, Francis Strobbe<sup>2</sup>, Andrew Skowno<sup>3</sup>, Vernon Visser<sup>1</sup>, Michael Brooks<sup>4</sup>, Yvan Stojanov<sup>2</sup>, Doug Harebottle<sup>5</sup>, Nancy Job<sup>3</sup>

<sup>1</sup> Centre for Statistics in Ecology, the Environment and Conservation, University of Cape Town, Cape Town, South Africa

<sup>2</sup> Operational Directorate Natural Environment, Royal Belgian Institute of Natural Sciences, Brussels, Belgium

<sup>3</sup> South African National Biodiversity Institute, Kirstenbosch Research Centre, Cape Town, South Africa

<sup>4</sup> FitzPatrick Institute of African Ornithology, University of Cape Town, Cape Town, South Africa

<sup>5</sup> Risk and Vulnerability Science Centre, Sol Plaatje University, Kimberley, South Africa

Correspondence\*:

Francisco Cervantes

f.cervantesperalta@gmail.com

## 2 ABSTRACT

With increasing pressure on ecosystems, we have seen substantial efforts in the collection and management of ecological data, over the last decade. This is especially true for freshwater ecosystems, which are among the most impacted by human activity and yet they have been lagging behind in terms of data availability, compared to other terrestrial ecosystems. Despite improvements in data collection and accessibility, raw data still need to be analysed and interrogated to support conservation programmes and management decisions, a process that can be complex and time consuming. The South African Biodiversity Data Pipeline for Wetlands and Waterbirds (BIRDIE) aims to help fast and efficient information uptake, bridging the gap between raw ecological datasets and final users. BIRDIE is a full data pipeline that takes up raw data, and through a series of processing steps, estimates indicators related to abundance, distribution and diversity of waterbirds, keeping track of their associated uncertainty. It focuses on the assessment of waterbird populations in South Africa and uses two citizen-science bird monitoring datasets, namely: the African Bird Atlas Project and the Coordinated Waterbird Counts. In addition, a suite of environmental layers help contextualise waterbird population indicators, and link these to the ecological condition of the supporting wetlands. Data processing is conveniently organised in modules that can be run independently, and include tasks, such as: occupancy modelling, state-space modelling, and computation of indicators at multiple temporal and spatial scales. Both data and indicators are accessible to end users through an online portal and through web services. Envisioned users of BIRDIE include government officials, conservation managers, researchers and the general public. Acknowledging that conservation programmes run at multiple scales, from site management to international agreements, we have developed a granular framework in which

# BIRDIE: A data pipeline to inform wetland and waterbird conservation at multiple scales

Francisco Cervantes<sup>1,2\*</sup>, Res Altwegg<sup>1</sup>, Francis Strobbe<sup>3</sup>, Andrew Skowno<sup>2</sup>,  
Vernon Visser<sup>1</sup>, Michael Brooks<sup>4</sup>, Yvan Stojanov<sup>3</sup>, Douglas M. Harebottle<sup>5</sup>,  
Nancy Job<sup>2</sup>

<sup>1</sup> Centre for Statistics in Ecology, the Environment and Conservation, University of Cape Town, Cape Town, South Africa

<sup>2</sup> South African National Biodiversity Institute, Kirstenbosch Research Centre, Cape Town, South Africa

<sup>3</sup> Operational Directorate Natural Environment, Royal Belgian Institute of Natural Sciences, Brussels, Belgium

<sup>4</sup> FitzPatrick Institute of African Ornithology, University of Cape Town, Cape Town, South Africa

<sup>5</sup> Risk and Vulnerability Science Centre, Sol Plaatje University, Kimberley, South Africa

Correspondence\*:

Francisco Cervantes

f.cervantesperalta@gmail.com

## 2 ABSTRACT

Efforts to collect ecological data have intensified over the last decade. This is especially true for freshwater habitats, which are among the most impacted by human activity and yet lagging behind in terms of data availability. Now, to support conservation programmes and management decisions, these data need to be analysed and interpreted; a process that can be complex and time consuming. The South African Biodiversity Data Pipeline for Wetlands and Waterbirds (BIRDIE) aims to help fast and efficient information uptake, bridging the gap between raw ecological datasets and the information final users need. BIRDIE is a full data pipeline that takes up raw data, and estimates indicators related to abundance, distribution and diversity of waterbirds, while keeping track of their associated uncertainty. At present, most functionality focuses on the assessment of waterbird populations in South Africa using two citizen-science bird monitoring datasets, namely: the African Bird Atlas Project and the Coordinated Waterbird Counts. In addition, a suite of environmental layers help contextualise waterbird population indicators, and link these to the ecological condition of the supporting wetlands. In the future, we aim to develop more indicators specific to the ecological structure and function of wetlands. Data processing is conveniently organised in modules that can be run independently, and include tasks, such as: data cleaning, statistical analysis, and computation of indicators at multiple temporal and spatial scales. Both data and indicators are accessible to end users through an online portal and web services. Envisioned users of BIRDIE include government officials, conservation managers, researchers and the general public, all of whom have been engaged throughout the project. Acknowledging that conservation programmes run at multiple spatial and temporal scales, we have developed a granular framework in which waterbird population indicators are estimated at

24 waterbird population indicators are estimated at small scales, and then these are aggregated  
25 to compute similar indicators at broader scales. The online portal is designed to provide spatial  
26 and temporal visualisation of the indicators using maps and time series, to help contextualization.  
27 This paper describes the structure of the BIRDIE pipeline and the technical features underpinning  
28 its components.

29 **Keywords:** Biodiversity informatics, Citizen science, Data pipeline, Waterbirds, Wetlands, Species distribution, Species Abundance,  
30 Diversity

## INTRODUCTION

31 Freshwater ecosystems are among the most productive, biodiverse, and most efficient at capturing and  
32 storing carbon (Convention on Wetlands, 2021). Unfortunately, they are also among the most impacted  
33 by human activity (Convention on Wetlands, 2021; Skowno et al., 2019), and climate change will likely  
34 increasingly exacerbate the pressure on freshwater resources. This is particularly true for the African  
35 continent, home to some of the largest wetlands, which not only host a wealth of freshwater species, but  
36 are also key in supporting human communities (Stephenson et al., 2020). Such critical issues have fuelled  
37 unprecedented efforts to collect and mobilise freshwater biodiversity data (Dallas et al., 2021; Wetzel et al.,  
38 2015). While we must strive to keep monitoring programmes that deliver data funded and alive, it is clear  
39 that data on their own are not enough (MacFadyen et al., 2022). If we are to take effective action to stop  
40 ecosystem degradation, it is important that data are analysed to extract indicators that are meaningful for  
41 decision- and policy-making (Jetz et al., 2019; P. Stephenson et al., 2017).

42 Bridging the gap between biodiversity data sets and conservation action requires, not only collecting,  
43 but also analysing and interpreting the data (Pollock et al., 2020). Furthermore, with continuous data  
44 collection, we need to implement workflows that update indicators and support decisions in a timely fashion  
45 (MacFadyen et al., 2022; Yenni et al., 2019). Automated data pipelines allow us to keep datasets updated  
46 and free of errors (Yenni et al., 2019), make model-based forecasts, and evaluate previous forecasts in light  
47 of new data (White et al., 2019). These modern and automated data workflows require multidisciplinary  
48 skills in ecology, statistics, data science, and software development, but their end products should ideally  
49 be free, accessible and easy to interpret (P. Stephenson et al., 2017). It would also be desirable that they  
50 integrate multiple datasets and environmental layers to produce a holistic understanding of biodiversity  
51 structure and function (MacFadyen et al., 2022).

52 South Africa is leading the African continent in terms of biodiversity data availability (Barnard et al.,  
53 2017), with successful citizen-science programmes such as the Southern African Bird Atlas Project (Brooks  
54 et al., 2022), and biodiversity data platforms, such as the Biodiversity Advisor (SANBI, 2023) or the  
55 Freshwater Biodiversity Information System (FBIS, Dallas et al., 2021). In contrast, dashboards and tools  
56 that facilitate the timely uptake of information and unlock the utility of current data are still limited. Here,  
57 we describe a data pipeline that implements a workflow of wetland- and waterbird-related biodiversity  
58 data, the South African Biodiversity Data Pipeline for Wetlands and Waterbirds (BIRDIE). At its core,  
59 BIRDIE utilises two long-term citizen-science programmes that have collected waterbird data in South  
60 Africa for more than two decades, and are still active: the Southern African Bird Atlas Project (SABAP;  
61 Brooks et al., 2022) and the Coordinated Waterbird Counts (CWAC; FIAO, 2022). These datasets have well  
62 defined protocols supported by an established community of citizen scientists, and they provide information  
63 about: 1) bird species occurrence, with visits to a grid of pentads (5' x 5' grid cells) across Southern Africa  
64 (SABAP), and 2) bird abundance, with waterbird counts taken twice a year at ca. 730 sites across South  
65 Africa (CWAC). These data provide the necessary information to investigate changes in species abundance

24 small scales, and then these are aggregated to compute similar indicators at broader scales.  
25 The online portal is designed to provide spatial and temporal visualisation of the indicators using  
26 maps, time series and pre-compiled reports for species, sites and conservation programmes.  
27 This paper describes the structure of the BIRDIE pipeline and the technical features underpinning  
28 its components.

29 **Keywords:** Biodiversity informatics, Citizen science, Data pipeline, Waterbirds, Wetlands, Species distribution, Species Abundance,  
30 Diversity

## INTRODUCTION

31 Freshwater ecosystems are among the most productive, biodiverse, and efficient at capturing and storing  
32 carbon (Convention on Wetlands, 2021). Unfortunately, they are also among the most impacted by human  
33 activity (Convention on Wetlands, 2021; Skowno et al., 2019), and climate change will likely exacerbate  
34 the pressure on freshwater resources. This is particularly true for the African continent, home to some of  
35 the largest wetlands, which not only host a wealth of freshwater species, but are also key in supporting  
36 human communities (Stephenson et al., 2020). Such critical issues have fuelled unprecedented efforts to  
37 collect and mobilise freshwater biodiversity data (Dallas et al., 2021; Wetzel et al., 2015).

38 While we must strive to keep monitoring programmes that deliver data funded and alive, it is clear that  
39 data on their own are not enough (MacFadyen et al., 2022). If we are to take effective action to stop  
40 ecosystem degradation, it is important that data are analysed to extract indicators that are meaningful for  
41 decision- and policy-making (Harebottle and Underhill 2016, Jetz et al., 2019; Stephenson et al., 2017).  
42 Furthermore, with continuous data collection, we need to implement workflows that update indicators  
43 and support decisions in a timely fashion (MacFadyen et al., 2022; Yenni et al., 2019). Automated data  
44 pipelines allow us to keep datasets updated and free of errors (Yenni et al., 2019), make model-based  
45 forecasts, and evaluate previous forecasts in light of new data (White et al., 2019). These modern and  
46 automated data workflows require multidisciplinary skills in ecology, statistics, data science, and software  
47 development, but their end products should ideally be free, accessible and easy to interpret (Stephenson  
48 et al., 2017). It would also be desirable that they integrate multiple datasets and environmental layers to  
49 produce a holistic understanding of biodiversity structure and function (MacFadyen et al., 2022).

50 South Africa is leading the African continent in terms of biodiversity data availability (Barnard et al.,  
51 2017), with successful citizen-science programmes such as the Southern African Bird Atlas Project (Brooks  
52 et al., 2022), and biodiversity data platforms, such as the Biodiversity Advisor (SANBI, 2023) or the  
53 Freshwater Biodiversity Information System (FBIS, Dallas et al., 2021). In contrast, dashboards and tools  
54 that facilitate the timely uptake of information and unlock the utility of current data are still limited. There  
55 is also an imbalance in data availability across taxonomic groups and habitats. Regular monitoring of the  
56 status, distribution and condition of wetlands ecosystems is urgently required to understand environmental  
57 pressures on wetland habitats, but challenges associated with limited human and budget capacity hamper  
58 the collection of the necessary data. Conversely, available waterbird species data are rich in detail and  
59 coverage, and could provide a stronger basis for both adaptive management and reporting at priority  
60 wetland sites.

61 Here, we describe a data pipeline that implements a workflow of wetland- and waterbird-related  
62 biodiversity data, the South African Biodiversity Data Pipeline for Wetlands and Waterbirds (BIRDIE). At  
63 present, most of BIRDIE's functionality focuses on computing indicators related to waterbird distribution  
64 and abundance, which are considered the minimum set of variables necessary to study changes in

66 and distribution, which are considered the minimum set of variables necessary to study changes in species  
67 populations (Pereira et al. 2013, Jetz et al. 2020). Apart from waterbird data, BIRDIE uses and serves  
68 ancillary environmental data for contextualising the aforementioned variables associated with waterbird  
69 populations, and also for describing the state of the wetlands that support them.

70 The BIRDIE data pipeline was designed to run periodically (yearly in principle), and automatically  
71 (but supervised), to keep information updated. It is meant to inform the general public about the state of  
72 wetlands and waterbird populations, but also to support the work of particular end users. A first group  
73 of users includes authorities that need to report on the state of wetlands (e.g., Ramsar Convention), or  
74 migratory waterbirds (e.g. African-Eurasian Migratory Waterbird Agreement) as required by international  
75 agreements, for national level processes (e.g., National Biodiversity Assessment) and for provincial or  
76 municipal level regulation (e.g. to set hunting quotas for waterfowl). A second group of users includes site  
77 managers and other stakeholders who need to make a range of decisions specific to certain wetlands (e.g.,  
78 related to wetland condition and extent) and the waterbirds supported by these habitats. Such policy-linked  
79 outputs are used to inform environmental strategies, identify priorities for protection and for sustainable  
80 use of biodiversity, and guide land-use management.

## FRAMEWORK AND TARGET USERS

81 Indicators on the state of biodiversity have been adopted by a range of multilateral environmental  
82 agreements including the United Nations Convention on Biological Diversity (CBD, 2022) and Sustainable  
83 Development Goals (SDGs; UN, 2022). New indicators are under development and established processes,  
84 such as the International Union for the Conservation of Nature (IUCN, 2022) species red-listing efforts,  
85 are receiving renewed attention (Han et al., 2017). With these indicators come various global and national  
86 initiatives and targets for reducing rates of biodiversity loss (Mace et al., 2018). Essential Biodiversity  
87 Variables (EBVs) have been conceptualised and developed to help standardise and improve interoperability  
88 of biodiversity data for monitoring (Pereira et al., 2013).

89 Within this framework, BIRDIE gives support to both national and international programs contributing  
90 information about the state of waterbird populations in South Africa, with a view to expand to the Southern  
91 Africa region. We focus primarily on species population EBVs, with the assessment of waterbird abundance,  
92 distribution and diversity, and changes of these over time (Jetz et al., 2019; Kissling et al., 2018). Species  
93 diversity falls under the community composition EBVs rather than species populations.

94 At the international scale, South Africa is signatory to the Ramsar Convention (Convention on Wetlands,  
95 2021), hosts 28 Wetlands of International Importance, and needs to produce reports on the state of sites  
96 every three years. Change in condition and extent of wetland habitat, as well as those Ramsar Criteria  
97 invoked when listing the different sites as internationally important are all core reporting requirements.  
98 These include changes in overall abundance and distribution of waterbirds at these sites, with special  
99 attention to threatened and migratory species.

100 National reports must also be compiled for the Agreement on the Conservation of African-Eurasian  
101 Migratory Waterbirds (AEWA; UNEP, 2022), another international agreement, framed under the Convention  
102 on Migratory Species, and focused on protecting migratory waterbirds and their habitats. Reporting for this  
103 agreement requires the full suite of EBVs outlined earlier in this section, now specific to migratory species,  
104 namely, abundance, distribution, diversity, and changes in these over time. Within South Africa, there is  
105 alignment and cooperation between Ramsar and AEWA, and as such, reporting on the change in wetland  
106 extent and condition is also relevant.

65 species populations (Pereira et al. 2013, Jetz et al. 2020). BIRDIE utilises two long-term citizen-science  
66 programmes that have collected waterbird data in South Africa for more than two decades, and are  
67 still active: the Southern African Bird Atlas Project (SABAP; Brooks et al., 2022) and the Coordinated  
68 Waterbird Counts (CWAC; FIAO, 2022). Apart from waterbird data, BIRDIE uses and serves ancillary  
69 environmental data for contextualising the aforementioned waterbird population variables, and also for  
70 describing the state of the wetlands that support them. In a next phase, we plan to expand the functionality  
71 of the pipeline to provide indicators of wetland ecosystem structure and function.

72 BIRDIE is embedded into the South African National Biodiversity Institute (SANBI) biodiversity  
73 informatics infrastructure and it was conceived as a tool to inform environmental strategies, identify  
74 priorities for the protection and sustainable use of biodiversity, and to guide land-use management. Because  
75 such policy-linked objectives require updated and timely information, the pipeline was designed to run  
76 periodically (yearly in principle), and automatically (but supervised). Currently, BIRDIE provides indicators  
77 for South Africa only, but in the future we expect to expand its coverage to other African countries. In what  
78 follows we describe BIRDIE's data pipeline workflow from data acquisition to display of final outputs, as  
79 well as the technologies we have used and the general modelling frameworks adopted.

## FRAMEWORK AND TARGET USERS

80 The main objective of BIRDIE is to provide information to support authorities that need to report on the  
81 state of wetlands or waterbird populations at multiple levels: 1) as required by national and international  
82 programmes and agreements, 2) provincial authorities, site managers and other stakeholders who need  
83 to make a range of decisions specific to certain wetlands, and 3) the general public could make use of  
84 BIRDIE's freely available outputs for a variety of reasons, including recreation and local conservation  
85 initiatives.

86 Indicators on the state of biodiversity have been adopted by a range of multilateral environmental  
87 agreements including the United Nations Convention on Biological Diversity (CBD, 2022) and Sustainable  
88 Development Goals (SDGs; UN, 2022). New indicators are under development and established processes,  
89 such as the International Union for the Conservation of Nature (IUCN, 2022) species red-listing efforts,  
90 are receiving renewed attention (Han et al., 2017). With these indicators come various global and national  
91 initiatives and targets for reducing rates of biodiversity loss (Mace et al., 2018). Essential Biodiversity  
92 Variables (EBVs) have been conceptualised and developed to help standardise and improve interoperability  
93 of biodiversity data and monitoring (Pereira et al., 2013). Within this framework, BIRDIE gives support to  
94 both national and international programs contributing information about the state of waterbird populations  
95 in South Africa, with a view to expand to the Southern Africa region. We focus primarily on species  
96 population EBVs, with the assessment of waterbird abundance, distribution and diversity, and changes of  
97 these over time (Jetz et al., 2019; Kissling et al., 2018).

98 At an international scale, the BIRDIE team has engaged in conversation with two strategic partners from  
99 the project outset: the Ramsar Convention Secretariat and the Technical Committee of the Agreement  
100 on the Conservation of African-Eurasian Migratory Waterbirds. South Africa is signatory to the Ramsar  
101 Convention (Convention on Wetlands, 2021), hosts 28 Wetlands of International Importance, and needs to  
102 produce reports on the state of these sites every three years. National reports must also be compiled for  
103 the Agreement on the Conservation of African-Eurasian Migratory Waterbirds (AEWA; UNEP, 2022), an  
104 international agreement, framed under the Convention on Migratory Species, and focused on protecting  
105 migratory waterbirds and their habitats. The Ramsar Convention and AEWA both require information on  
106 changes in overall abundance and distribution of waterbirds, with AEWA focusing on migratory species.

107 At the national level, South Africa produces a National Biodiversity Assessment every four years, which  
108 constitutes the main reporting tool of the state of biodiversity in the country, and informs policy and  
109 conservation strategies (Skowno et al., 2019). At the same time, there are regular efforts to address the  
110 conservation status of South African species within the IUCN Red-List framework. Changes in abundance  
111 and distribution of species are key in these assessments to track and report on population trends, and shifts  
112 in species ranges and community diversity.

113 Keeping these main reporting channels in mind, BIRDIE also intends to support local management  
114 actions and basic research. Site-scale wetland monitoring is severely limited in South Africa, lagging far  
115 behind monitoring of other aquatic ecosystems such as rivers and estuaries. Managers ideally need to report  
116 on the state of the wetland (e.g., wetland condition, flux in surface water extent) as well as the species  
117 that the wetland supports, including species of special concern. Local waterbird and wetland information  
118 can facilitate the development of site-specific management actions and management plans, and support  
119 permitting decisions. At the same time, linking the local manager inputs and feedback into the data pipeline  
120 closes the gap between large-scale assessments and local data collection. The data pipeline also allows  
121 citizen scientists to more actively interact with the data they have collected, and to see it taken up into the  
122 statistical analyses and data visualisations.

## INPUT DATA

123 South Africa needs to monitor its biodiversity to inform decisions that affect the environment, and to fulfil  
124 reporting obligations linked to international conventions. Fortunately, we have a number of long-running  
125 citizen science projects that help monitor waterbird populations throughout the country. At its core, the  
126 project leverages two bird-related datasets: the Coordinated Waterbird Counts (CWAC) and the South  
127 African Bird Atlas Project (SABAP2), which is part of the larger African Bird Atlas Project (ABAP). These  
128 datasets have well established citizen scientist support and offer information about: 1) bird abundance, with  
129 waterbird counts taken twice a year at 731 water bodies across Southern Africa (mostly South Africa) since  
130 1992, and 2) species occurrence, with visits to a grid of pentads (5' x 5' grid cells) initiated in 2007 and  
131 covering several African countries.

132 The Coordinated Waterbird Counts project provides regular counts of all waterbirds at just over 700  
133 sites throughout South Africa. The project was launched in 1992 and since then, it has accumulated a long  
134 time series for many sites. However, not all sites have been monitored since the start of the project, some  
135 regions are better represented than others, and not all sites have been monitored continuously (Figure 3).  
136 Waterbird species have diverse habitat requirements and life histories; some use the same sites year-round,  
137 whereas others are migratory. To capture this diversity, CWAC counts are carried out twice per year: one in  
138 mid-summer and one in mid-winter. These data are noisy because it is difficult to count waterbirds precisely,  
139 but with appropriate statistical analyses, they can reveal long-term temporal trends and seasonal fluctuations  
140 in waterbird populations. Waterbird population sizes and trends are important reporting parameters for  
141 international agreements such as the Ramsar Convention, or AEWA.

142 ABAP offers occurrence, rather than abundance data. In ABAP, volunteers collect checklists of all birds  
143 observed over a grid of pentads (5' x 5' minute grid) covering different African countries. We are currently  
144 restricting our analysis to South Africa, and therefore we are using the SABAP2 component of ABAP  
145 (Figure 4). However, in the future we would like to expand our functionality to cover other countries  
146 contributing data to ABAP, such as Kenya or Nigeria. Under the SABAP2 protocol, which started in 2007,  
147 observers need to spend at least two hours of intensive birding at a pentad and are asked to visit as many  
148 habitats within it as possible. They can add new species for up to five days. SABAP2 currently has ca. 17

107 Both conventions also report on indicators such as change in wetland extent and condition. Engagement  
108 with the South African national government bodies for both of these conventions ensures the reporting  
109 component of the BIRDIE project responds directly to their needs.

110 At the national level, South Africa produces a National Biodiversity Assessment every four years, which  
111 constitutes the main reporting tool of the state of biodiversity in the country, and informs policy and  
112 conservation strategies (Skowno et al., 2019). At the same time, there are regular efforts to address the  
113 conservation status of South African species within the IUCN Red-List framework. Changes in abundance  
114 and distribution of species are key in these assessments to track and report on population trends, and  
115 shifts in species ranges and community diversity. BIRDIE is embedded within SANBI, which is the  
116 organisation mandated to report on the state of biodiversity in South Africa. As such, the outputs produced  
117 by the pipeline have a direct connection to needs specified for National Biodiversity Assessments, the  
118 Freshwater Biodiversity Programme and other national decision processes regarding freshwater ecosystems  
119 and species.

120 Keeping these main reporting channels in mind, BIRDIE also intends to support local management  
121 actions and basic research. Site-scale wetland monitoring is severely limited in South Africa, lagging far  
122 behind monitoring of other aquatic ecosystems such as rivers and estuaries. Managers ideally need to report  
123 on the state of the wetland (e.g., wetland condition, flux in surface water extent) as well as the species  
124 that the wetland supports, including species of special concern. Local waterbird and wetland information  
125 can facilitate the development of site-specific management actions and management plans, and support  
126 permitting decisions. At the same time, linking the local manager inputs and feedback into the data pipeline  
127 closes the gap between large-scale assessments and local data collection. In this sense, throughout the  
128 development of the pipeline, we have engaged with stakeholders at a pilot site, the Barberspan Nature  
129 Reserve. These conversations were enormously insightful to understand the variety of questions that may  
130 arise when working at a local level. One key take-away message from these engagements was that we  
131 should favour a flexible online portal, where users can customise their queries, over a rich but fixed set of  
132 outputs.

133 Finally, we hope that the data pipeline will also allow citizen scientists to more actively interact with the  
134 data they have collected, and to see it taken up into the statistical analyses and data visualisations. The  
135 general public could also benefit from a flexible wetland and waterbird portal, with the right information to  
136 aid their interpretation.

## INPUT DATA

137 In South Africa, we have a number of long-running citizen science projects that help monitor waterbird  
138 populations throughout the country. At its core, BIRDIE leverages two bird-related datasets: the Coordinated  
139 Waterbird Counts (CWAC, FIAO, 2022) and the second phase of the South African Bird Atlas Project  
140 (SABAP2, Brooks et al., 2022), which is part of the larger African Bird Atlas Project (ABAP). These  
141 datasets have well established citizen scientist support and offer information about: 1) bird abundance, with  
142 waterbird counts taken twice a year at 731 water bodies across Southern Africa (mostly South Africa) since  
143 1992, and 2) species occurrence, with visits to a grid of pentads (5' x 5' grid cells) initiated in 2007 and  
144 covering several African countries.

145 The Coordinated Waterbird Counts project provides regular counts of all waterbirds at just over 700 sites  
146 throughout South Africa. Counts are predominantly conducted by field observers from a set of observation  
147 points defined for each site, and that are visited twice a year; although in some sites other types of counts,  
148 such as count by boat, are also used (FIAO, 2022). The project was launched in 1992 and since then, it

149 million records, and > 2 million records are added per year. The structured sampling protocol, together  
150 with the spatial and temporal extent of SABAP2 allow us to examine how bird distributions are changing  
151 over time, although statistical modelling is required to account for imperfect detection and sampling biases  
152 (Figure 4).

153 There are a variety of other data sources that BIRDIE uses for adding environmental information into its  
154 analytical workflows. Most of these data sources are conveniently accessed through Google Earth Engine,  
155 such as TerraClimate (Abatzoglou et al., 2018), the JRC surface water dataset (Pekel et al., 2016), MODIS  
156 Vegetation Indices (Didan, Kamel, 2015) and Digital Elevation Models (DEM, (Yamazaki et al., 2017)).  
157 Other data not yet available on Google Earth Engine, such as the National Wetland Map (SANBI, in prep.)  
158 are managed independently.

## INDICATORS AND STATISTICAL METHODS

159 Capturing good quality raw data is a fundamental first step to monitor the state of biodiversity. However,  
160 capturing error-free comprehensive datasets over large temporal and spatial scales requires huge efforts  
161 that are unrealistic in most situations. In fact, the very nature of natural systems often prevents them  
162 from being observed perfectly and entirely. Therefore, some level of statistical analysis is required to  
163 estimate the state of the system of interest, and separate it from observational artefacts introduced by the  
164 observation process used for capturing the data (Gimenez et al., 2014; King, 2014; Yoccoz et al., 2001).  
165 The BIRDIE pipeline broadly uses two types of models: i) occupancy models (Altwegg and Nichols, 2019;  
166 MacKenzie et al., 2002) to estimate the probability of species being present at the different SABAP2  
167 pentads, and ii) state-space models (Buckland et al., 2004; Newman et al., 2014) to estimate the number of  
168 individuals at the sites monitored by the CWAC programme. Contrary to raw observations (counts and  
169 detection/non-detection of a species), model-based estimates (abundance and occupancy probabilities)  
170 allow us to quantify uncertainty.

171 The variety of end-user needs requires a pipeline that provides waterbird population indicators at multiple  
172 spatial and temporal scales. Therefore, in addition to estimating basic occupancy and abundance at small  
173 scales (i.e., individual site/pentad), the BIRDIE pipeline produces other high-level indicators obtained  
174 by aggregation (Table 1). The idea is to follow a process whereby raw data are used to inform models  
175 that estimate indicators at the smallest temporal and spatial scales possible, and then to aggregate these  
176 estimates at larger scales, as required. For example, species abundance can be estimated for a set of  
177 regularly monitored wetlands in South Africa, and these site-specific estimates can then be combined to  
178 calculate an abundance index for all sites as a group. We can follow this procedure to estimate abundance  
179 and occupancy probabilities at national, regional and local levels, as well as for specific groups of wetlands  
180 (e.g., designated Ramsar sites, estuaries or artificial sites (e.g., dams)).

181 The main indicators computed by the BIRDIE pipeline for waterbird species are:

- 182 • Population size: estimated for CWAC sites in two seasons per year. For each species, only those  
183 wetlands with at least a ten-year coverage between 1997 and 2021 are analysed.
- 184 • Occurrence: estimated for ABAP pentads on an annual basis.
- 185 • Diversity: the simplest and most easily understood metric is species richness. Species richness can  
186 be calculated based on the occupancy analysis, by summing occupancy probabilities of all species  
187 potentially present in each pentad, to estimate the expected number of species present.
- 188 • Important records: sightings of rarities, invasive species. Although this information does not require  
189 any statistical processing, it does make particular records more visible.

149 has accumulated a long time series for many sites. However, not all sites have been monitored since the  
150 start of the project, some regions are better represented than others, and not all sites have been monitored  
151 continuously (Figure 2). Waterbird species have diverse habitat requirements and life histories; some use the  
152 same sites year-round, whereas others are migratory or undergo local movements. To capture this diversity,  
153 CWAC counts are carried out twice per year: once in mid-summer and once in mid-winter. Although counts  
154 incorporate errors due to imperfect waterbird detection by observers, with appropriate statistical analyses,  
155 they can reveal long-term temporal trends and seasonal fluctuations in waterbird populations.

156 ABAP offers occurrence, rather than abundance data. In ABAP, volunteers collect checklists of all birds  
157 observed over a grid of pentads (5' x 5' minute grid) covering different African countries (Brooks et al.,  
158 2022). We are currently restricting our analysis to South Africa, and therefore we are using the SABAP2  
159 component of ABAP (Figure 3). However, we plan to expand BIRDIE's functionality to cover other  
160 countries contributing data to ABAP, such as Kenya or Nigeria. Under the SABAP2 protocol, which started  
161 in 2007, observers need to spend at least two hours of intensive birding at a pentad and are asked to visit  
162 as many habitats within it as possible. They can add new species for up to five days. SABAP2 currently  
163 has ca. 17 million records, and > 2 million records are added per year. The structured sampling protocol,  
164 together with the spatial and temporal extent of SABAP2 allow us to examine how bird distributions are  
165 changing over time, although statistical modelling is required to account for imperfect detection and spatial  
166 sampling biases (Figure 3).

167 There are a variety of other data sources that BIRDIE uses for adding environmental information into its  
168 analytical workflows. Most of these data sources are conveniently accessed through Google Earth Engine,  
169 such as TerraClimate (Abatzoglou et al., 2018), the JRC surface water dataset (Pekel et al., 2016), MODIS  
170 Vegetation Indices (Didan, Kamel, 2015) and Digital Elevation Models (DEM, Yamazaki et al., 2017).  
171 Other data not yet available on Google Earth Engine, such as the National Wetland Map (SANBI, in prep.)  
172 are managed independently.

## INDICATORS AND STATISTICAL METHODS

173 Capturing good quality raw data is a fundamental first step to monitor the state of biodiversity. However,  
174 raw data reflect not only the biological signal of interest but also the sampling process, which is typically  
175 spatially biased and subject to imperfect detection (Yoccoz et al., 2001). Therefore, some level of statistical  
176 analysis is required to estimate the state of the system of interest, and separate it from observational  
177 artefacts introduced by the observation process used for capturing the data (Gimenez et al., 2014; King,  
178 2014; Yoccoz et al., 2001). The BIRDIE pipeline broadly uses two types of models: 1) occupancy models  
179 (Altwegg and Nichols, 2019; MacKenzie et al., 2002) to estimate the probability of a species being present  
180 at the different SABAP2 pentads, and 2) state-space models (Buckland et al., 2004; Newman et al., 2014)  
181 to estimate the number of individuals at the sites monitored by the CWAC programme. Contrary to raw  
182 observations (counts and detection/non-detection of a species), model-based estimates (abundance and  
183 occupancy probabilities) allow us to quantify uncertainty.

184 The variety of end-user needs requires a pipeline that provides waterbird population indicators at multiple  
185 spatial and temporal scales. Therefore, in addition to estimating basic occupancy and abundance at small  
186 scales (i.e., individual site/pentad), the BIRDIE pipeline produces other high-level indicators obtained  
187 by aggregation (Table 1). The idea is to follow a process whereby raw data are used to inform models  
188 that estimate indicators at the smallest temporal and spatial scales possible, and then to aggregate these  
189 estimates at larger scales, as required. For example, species abundance can be estimated for a set of  
190 regularly monitored wetlands in South Africa, and these site-specific estimates can then be combined to

190 In addition to estimates of static indicators, the pipeline also estimates their associated dynamics, such as:  
 191 changes in abundance, occupancy probabilities and diversity. The temporal reference for these dynamics  
 192 can also vary ranging from a single season to multiple years (typically ca. 5 years, for short-term changes,  
 193 and ca. 15 years for long-term changes). It is important that uncertainty is correctly propagated when  
 194 aggregating, and also when estimating dynamic indicators. We work in a Bayesian framework and use the  
 195 posterior distribution of occupancy probabilities and species abundance to define indicators at the various  
 196 scales. Working with full posterior distributions allows us to conveniently keep track of the uncertainty in  
 197 the estimates used as building blocks for other derived indicators.

### 198 Delineating species distributions

199 Occupancy models are fitted to detection/non-detection data from SABAP2 to delineate the distribution of  
 200 waterbird species and its dynamics over time. Within the SABAP2 framework, observers visit pentads and  
 201 make a list of the bird species detected during the visit. We assume that observers identify species correctly  
 202 and only list species observed (the rigorous vetting process of SABAP2 data justifies this assumption), but  
 203 non-detections may be caused by either species not being present in the pentad or by observers not being  
 204 able to detect them, when present. Therefore, occupancy models describe two processes simultaneously: i)  
 205 the underlying occupancy of the sites (pentads), and ii) the observation process whereby species present  
 206 might or might not be observed.

207 More precisely, we define  $z_{jt}$  to be the true occupancy of site  $j$  in year  $t$ , which can be 1 (if species  
 208 present) or 0 (if species absent) and has distribution:

$$z_{jt} | \psi_{jt} \sim \text{Bernoulli}(\psi_{jt})$$

209 where  $\psi_{jt}$  is the occupancy probability at site  $j$  and year  $t$ . The logit transformation of  $\psi_{jt}$  can be  
 210 modelled as a linear combination of covariates and smooth functions of covariates, such that:

$$\text{logit}(\psi_{jt}) = \mathbf{x}_{jt}^\top \boldsymbol{\beta} + \sum_{k=1}^K f_k(u_{jk})$$

211 where  $u_{jk}$  is a smooth function of the covariate  $u_k$ , which is defined as

$$f_k(u_{jk}) = \sum_{l=1}^L \mathbf{B}_l(u_{jkl}) \gamma_{jkl}$$

212 where the smooth function  $f$  is represented by a set of  $L$  basis functions  $\mathbf{B}_l$  evaluated at the value of the  
 213 covariates associated with site  $j$  at year  $t$ .

214 Then, the probability of detection of a species that is present in site  $j$  on visit  $i$  is denoted by  $p_{ij}$ . Following  
 215 the same logic as for the probability of occupancy, the logit transformation of  $p$  is modelled as a linear  
 216 combination of covariates and smooth functions:

$$\text{logit}(p_{ij}) = \mathbf{w}_{ij}^\top \boldsymbol{\alpha} + \sum_{h=1}^H f_h(v_{ih})$$

217 ,

191 calculate an abundance index for all sites as a group. We can follow this procedure to estimate abundance  
 192 and occupancy probabilities at national, regional and local levels, as well as for specific groups of wetlands  
 193 (e.g., designated Ramsar sites, estuaries or artificial sites).

194 The main indicators computed by the BIRDIE pipeline for waterbird species are:

- 195 • Abundance: estimated for CWAC sites in two seasons per year. For each species, only those wetlands  
 196 with at least a ten-year coverage between 1997 and 2021 are analysed statistically.  
 197 • Occurrence: estimated for ABAP pentads on an annual basis.  
 198 • Diversity: the simplest and most easily understood metric is species richness. Species richness can  
 199 be calculated based on the occupancy analysis, by summing occupancy probabilities of all species  
 200 potentially present in each pentad, to estimate the expected number of species present.  
 201 • Important records: sightings of rarities, invasive species. Although this information does not require  
 202 any statistical processing, it does make particular records more visible.

203 In addition to estimates of static indicators, the pipeline also estimates their associated dynamics, such as:  
 204 changes in abundance, occupancy probabilities and diversity. The temporal reference for these dynamics  
 205 can also vary ranging from a single season to multiple years (typically ca. 5 years, for short-term changes,  
 206 and ca. 15 years for long-term changes).

207 It is important that uncertainty is correctly propagated when aggregating, and also when estimating  
 208 dynamic indicators. We work in a Bayesian framework and use the posterior distribution of occupancy  
 209 probabilities and species abundance to define indicators at the various scales. Working with full posterior  
 210 distributions allows us to conveniently keep track of the uncertainty in the estimates used as building blocks  
 211 for other derived indicators.

## 212 Delineating species distributions

213 Occupancy models are fitted to detection/non-detection data from SABAP2 to delineate the distribution of  
 214 waterbird species and its dynamics over time. Within the SABAP2 framework, observers visit pentads and  
 215 make a list of the bird species detected during the visit. We assume that observers identify species correctly  
 216 and only list species observed (the rigorous vetting process of SABAP2 data justifies this assumption), but  
 217 non-detections may be caused by either species not being present in the pentad or by observers not being  
 218 able to detect them, when present. Therefore, occupancy models describe two processes simultaneously: i)  
 219 the underlying occupancy of the sites (pentads), and ii) the observation process whereby species present  
 220 might or might not be observed.

221 More precisely, we define  $z_{jt}$  to be the true occupancy of site  $j$  in year  $t$ , which can be 1 (if species  
 222 present) or 0 (if species absent) and has distribution:

$$z_{jt} | \psi_{jt} \sim \text{Bernoulli}(\psi_{jt})$$

223 where  $\psi_{jt}$  is the occupancy probability at site  $j$  and year  $t$ . The logit transformation of  $\psi_{jt}$  can be  
 224 modelled as a linear combination of covariates and smooth functions of covariates, such that:

$$\text{logit}(\psi_{jt}) = \mathbf{x}_{jt}^\top \boldsymbol{\beta} + \sum_{k=1}^K f_k(u_{jk})$$

218 Finally, the likelihood of observation  $y_{ij}$  can be written as:

$$y_{ij}|z_{jt}, p_{ij} \sim \text{Bernoulli}(z_{jt}p_{ij})$$

219 Spatial, spatio-temporal, and unstructured random effects can be specified for either occupancy or  
220 detection probabilities to account for variation across sites, observers and visits, not accounted for by the  
221 covariates incorporated in the models.

222 Each checklist is treated as an independent survey, but occupancy is assessed on a yearly basis. This  
223 means that if a species is detected in any one survey it is considered present that year. Therefore, missing a  
224 species because it left the site is considered part of the observation process and not the occupancy process.  
225 Migratory birds, for example, are considered present at a site even if they are only there for part of the year.

226 We are fitting single-season occupancy models without spatial random effects to most species. However,  
227 all models incorporate random effects to account for pentad- and observer-specific detection probabilities.  
228 If model diagnostics indicate poor model fit (see model diagnostics section below), we try incorporating  
229 spatial random effects for occupancy probabilities with an exponential decay function. Currently, we fit  
230 the models in R (R Core Team, 2022), in a Bayesian framework using the package *spOccupancy* (Doser  
231 et al., 2022), and running three MCMC chains for 20,000 iterations, with a thinning interval of 20. We  
232 use non-informative priors for all parameters when no information from other years is available, but we  
233 incorporate information obtained from other model fits if available, by centering the priors on the closest  
234 model's posterior means. However, it is important noticing that modelling details may differ among species  
235 and may be updated in the future.

### 236 Estimating abundance and population trends

237 State-space models (Buckland et al., 2004; Newman et al., 2014) are used to describe and understand  
238 dynamic systems that may not be perfectly observed. Within this framework, we consider waterbird  
239 abundance to be a process that evolves over time, and which we observe during visits to CWAC sites.  
240 However, counts conducted by observers are distorted by imperfect detection that translates into counting  
241 errors. By counting repeatedly over time, and assuming that abundance evolves smoothly over time  
242 compared to observation error, we can disentangle these two processes.

243 We consider that the observed counts ( $y_i$ ) at sampling occasion  $i$  (generally there were two sampling  
244 occasions per year, one in mid-summer and one in mid-winter), at any given site, arise from a  $\text{Poisson}(\lambda_i)$   
245 distribution

$$y_i \sim \text{Poisson}(\lambda_i)$$

246 And we model the log of the intensity  $\lambda_i$  as:

$$\log(\lambda_i) \sim N(\mu_i, \sigma^2)$$

247 where  $\mu_i$  is the mean abundance of waterbirds present at a site on sampling occasion  $i$  and  $\sigma^2$  is the  
248 corresponding variance of the observers counting error, both in the log scale. Therefore, counts depend  
249 both on the number of waterbirds present on site, and on errors in the counts of these birds.

250 To model changes in waterbird abundance between the two-seasons of year  $t$ , we define  $s_t$  to be the  
251 summer abundance and  $w_t$  the winter abundance. Note that there might be multiple counts in a single year

225 where  $u_{jk}$  is a smooth function of the covariate  $u_k$ , which is defined as

$$f_k(u_{jk}) = \sum_{l=1}^L \mathbf{B}_l(u_{jkl})\gamma_{jkl}$$

226 where the smooth function  $f$  is represented by a set of  $L$  basis functions  $\mathbf{B}_l$  evaluated at the value of the  
227 covariates associated with site  $j$  at year  $t$ .

228 We can then write the likelihood of observation  $y_{ij}$  as:

$$y_{ij}|z_{jt}, p_{ij} \sim \text{Bernoulli}(z_{jt}p_{ij})$$

229 The probability of detection of a species that is present in site  $j$  on visit  $i$  is denoted by  $p_{ij}$ . Following  
230 the same logic as for the probability of occupancy, the logit transformation of  $p$  is modelled as a linear  
231 combination of covariates and smooth functions:

$$\text{logit}(p_{ij}) = \mathbf{w}_{ij}^\top \boldsymbol{\alpha} + \sum_{h=1}^H f_h(v_{ih})$$

232 ,

233 Spatial, spatio-temporal, and unstructured random effects can be specified for either occupancy or  
234 detection probabilities to account for variation across sites, observers and visits, not accounted for by the  
235 covariates incorporated in the models.

236 Each checklist is treated as an independent survey, but occupancy is assessed on a yearly basis. This  
237 means that if a species is detected in any one survey it is considered present that year. Therefore, missing a  
238 species because it left the site is considered part of the observation process and not the occupancy process.  
239 Migratory birds, for example, are considered present at a site even if they are only there for part of the year.

240 We are fitting single-season occupancy models without spatial random effects to most species. However,  
241 all models incorporate random effects to account for pentad- and observer-specific detection probabilities.  
242 If model diagnostics indicate poor model fit (see model diagnostics section below), we assess models  
243 individually to understand the reasons, and if necessary we add spatial random effects for occupancy  
244 probabilities with an exponential decay function. Currently, we fit the models in R (R Core Team, 2022), in  
245 a Bayesian framework using the package *spOccupancy* (Doser et al., 2022), and running three MCMC  
246 chains for 20,000 iterations, with a thinning interval of 20. We use non-informative priors for all parameters  
247 when no information from other years is available, but we incorporate information obtained from other  
248 model fits if available, by centering the priors on the closest model's posterior means. However, it is  
249 important noticing that modelling details may differ among species and may be updated in future versions  
250 of BIRDIE.

## 251 Estimating abundance and population trends

252 State-space models (Buckland et al., 2004; Newman et al., 2014) are used to describe and understand  
253 dynamic systems that may not be perfectly observed . Within this framework, we consider waterbird  
254 abundance to be a process that evolves over time, and which we observe during visits to CWAC sites.  
255 However, counts conducted by observers are distorted by imperfect detection that translates into counting

252 and season, but the underlying true abundance is considered to stay constant in any given year and season  
 253 (for clarity, note also that while sampling occasions were indexed by  $i$ , years are indexed by  $t$ ). Thus, the  
 254 expected (log) abundance for any given count can be written as

$$\mu_i = s_t \text{summer} + w_t \text{winter}$$

255 where ‘summer’ is an indicator variable that takes on the value 1 in summer and 0 in winter, and ‘winter’  
 256 is the opposite. We then define abundance dynamics as:

$$s_t = s_{t-1} + \beta_t$$

$$w_t = s_t + \xi_t$$

258 where  $\beta_t$  corresponds to the change in summer abundance from year  $t-1$  to year  $t$ , and  $\xi_t$  is the difference  
 259 between summer and winter abundance, both in the log scale. If exponentiated, these parameters can  
 260 be interpreted as the rate of change in the population and the winter-to-summer ratio of the population,  
 261 respectively.

262 We impose relatively smooth changes in abundance by defining autocorrelation in  $\beta_t$  terms over time. In  
 263 addition, we define relationships between the rate of change in the population and environmental covariates.  
 264 These relationships facilitate the estimation of abundance for those years in which counts are missing, and  
 265 it is particularly useful to contain uncertainty in long periods with missing data between counts. Thus, we  
 266 set

267 We impose relatively smooth changes in abundance by defining autocorrelation in  $\beta_t$  and  $\xi_t$  terms over  
 268 time. In addition, we define relationships between the rate of change in the population  $\beta_t$  and environmental  
 269 covariates. These relationships facilitate the estimation of abundance for those years in which counts are  
 270 missing, and it is particularly useful to contain uncertainty in long periods with missing data between  
 271 counts. Thus, we set

$$\beta_t = \phi \beta_{t-1} + \eta_{t-1} + \zeta_{t-1}$$

$$\xi_t = \xi_{t-1} + \epsilon_{t-1}$$

273 where  $\phi$  lies between zero and one, and it defines an autoregressive term on  $\beta_{t-1}$ ;  $\eta_t$  captures the effect  
 274 of covariates in the expected change in abundance, and can be expanded to  $\gamma^T U$ , where  $U$  is a matrix of  
 275 covariate values and  $\gamma$  a vector of coefficients;  $\zeta_t$  and  $\epsilon_t$  are random variables that represent change in  
 276 abundance change, and change in winter to summer ratio, respectively.

277 We mentioned at the beginning that this model applies to each monitored site. However, we have multiple  
 278 sites, and counts are often missing for some seasons or even full years. To facilitate the estimation of  
 279 abundance with missing data, we borrow information from sites with counts, by defining a hierarchical  
 280 structure such that:

$$\zeta_{tj} \sim N(0, \sigma_{\zeta t}^2)$$

$$\epsilon_{tj} \sim N(0, \sigma_{\epsilon t}^2)$$

256 errors. By counting repeatedly over time, and assuming that abundance evolves smoothly over time  
 257 compared to observation error, we can disentangle these two processes.

258 We consider that the observed counts ( $y_i$ ) at sampling occasion  $i$  (generally there were two sampling  
 259 occasions per year, one in mid-summer and one in mid-winter), at any given site, arise from a Poisson( $\lambda_i$ )  
 260 distribution

$$y_i \sim \text{Poisson}(\lambda_i)$$

261 And we model the log of the intensity  $\lambda_i$  as:

$$\log(\lambda_i) \sim N(\mu_i, \sigma^2)$$

262 where  $\mu_i$  is the mean abundance of waterbirds present at a site on sampling occasion  $i$  and  $\sigma^2$  is the  
 263 corresponding variance of the observers counting error, both in the log scale. Therefore, counts depend  
 264 both on the number of waterbirds present on site, and on errors in the counts of these birds.

265 To model changes in waterbird abundance between the two-seasons of year  $t$ , we define  $s_t$  to be the  
 266 summer abundance and  $w_t$  the winter abundance. Note that there might be multiple counts in a single year  
 267 and season, but the underlying true abundance is considered to stay constant in any given year and season  
 268 (for clarity, note also that while sampling occasions were indexed by  $i$ , years are indexed by  $t$ ). Thus, the  
 269 expected (log) abundance for any given count can be written as

$$\mu_i = s_t \text{summer} + w_t \text{winter}$$

270 where ‘summer’ is an indicator variable that takes on the value 1 in summer and 0 in winter, and ‘winter’  
 271 is the opposite. We then define abundance dynamics as:

$$\begin{aligned} s_t &= s_{t-1} + \beta_t \\ w_t &= s_t + \xi_t \end{aligned}$$

273 where  $\beta_t$  corresponds to the change in summer abundance from year  $t-1$  to year  $t$ , and  $\xi_t$  is the difference  
 274 between summer and winter abundance, both in the log scale. If exponentiated, these parameters can  
 275 be interpreted as the rate of change in the population and the winter-to-summer ratio of the population,  
 276 respectively.

277 We impose relatively smooth changes in abundance by defining autocorrelation in  $\beta_t$  and  $\xi_t$  terms over  
 278 time. In addition, we define relationships between the rate of change in the population  $\beta_t$  and environmental  
 279 covariates. These relationships facilitate the estimation of abundance for those years in which counts are  
 280 missing, and it is particularly useful to contain uncertainty in long periods with missing data between  
 281 counts. Thus, we set

$$\beta_t = \phi \beta_{t-1} + \eta_{t-1} + \zeta_{t-1}$$

$$\xi_t = \xi_{t-1} + \epsilon_{t-1}$$

282 so that random changes at any site and year come from a common distribution of changes across all  
283 sites for that year. These distributions are normal with variances  $\sigma_{\zeta t}^2$  and  $\sigma_{et}^2$  for changes in abundance and  
284 winter to summer ratio, respectively.

285 We fit these models in R (R Core Team, 2022) with the additional functionality provided by JAGS  
286 (Plummer, 2003) using the package jagsUI (Kellner, 2021). We work on a Bayesian framework, using  
287 non-informative priors, and running three chains for 10,000 iterations each. Similar to the occupancy  
288 models, these are the details of the models we are working with at the moment, and they are intended  
289 to give an idea of the type of model we are using. These models might be updated in the future and the  
290 updated modelling details will be published on the BIRDIE website.

## 291 Data and model diagnostics

292 The pipeline needs to run for a multitude of species, with different ecological requirements and  
293 geographical distributions. Therefore, finding a model that suits all species is challenging. Not only  
294 may a model not be a good fit for a particular species, but the algorithms used for fitting the model may fail  
295 to converge due to characteristics of the data.

296 In a first control stage, we have defined some minimum requirements data should meet to enter the  
297 model-fitting process. Species that have been observed in five or less pentads in a year are considered to  
298 not have enough data to inform an occupancy model. Similarly, we chose only those CWAC sites that have  
299 been counted at least ten times between 1993 and 2021, to fit state-space models. Otherwise, data tend to  
300 be too sparse to assess trends in abundance reliably. These thresholds are rather arbitrary, based on our own  
301 preliminary experiences, and they are considered to be the minimum requirements for models to converge  
302 successfully. However, meeting these requirements does not guarantee model convergence or a good fit. To  
303 keep track of potential issues arising during model fitting, and to improve the algorithms of the pipeline,  
304 each time the pipeline runs it generates several reports that are later examined.

305 To decide whether any occupancy or state-space model converges, we calculate the Gelman-Rubin  
306 (Rhat) diagnostic (Gelman et al., 2014) for each estimated parameter. These diagnostics are then tabulated  
307 and stored for future revision. Any Rhat value above 1.1 or below 0.9 is considered to represent lack of  
308 convergence. Distinctive characteristics of the models with convergence issues are explored and addressed  
309 on a case by case basis, after the pipeline has finished running.

310 In addition to convergence, we assess goodness of fit using posterior predictive checks (Doser et al., 2022;  
311 Gelman et al., 2014). This procedure compares some quantity of interest calculated using pseudo-data  
312 simulated from the model posterior distribution, with that same quantity calculated from the observed data.  
313 In a well-fitting model we would expect real and synthetic data to produce similar values. For occupancy  
314 models, we produce simulated detection/non-detection data for each site, species and year and compute the  
315 expected number of detections out of as many visits as there were in the data. We compare the results of  
316 the simulations with the observed number of detections recorded in the data using a Chi-square test. For  
317 state-space models we follow a similar procedure, but instead of simulating detection/non-detection data  
318 for one year, we simulate count data for summer and winter, and aggregate these in a single annual count.  
319 Results from the goodness of fit Chi-square tests are also tabulated and stored for revision. Significant  
320 deviations detected with these tests are addressed for each case individually.

321 Due to the computational burden of the pipeline, it is not possible to run multiple models for each  
322 species, site and year, to perform model selection. Therefore, model selection is performed on a sample  
323 of species, selected to have representation of common and scarce taxa, but that are otherwise selected  
324 arbitrarily. Our general approach has been to include a rich set of variables that we believe can explain

283 where  $\phi$  lies between zero and one, and it defines an autoregressive term on  $\beta_{t-1}$ ;  $\eta_t$  captures the effect  
 284 of covariates in the expected change in abundance, and can be expanded to  $\gamma^\top U$ , where  $U$  is a matrix of  
 285 covariate values and  $\gamma$  a vector of coefficients;  $\zeta_t$  and  $\epsilon_t$  are random variables that represent change in  
 286 abundance change, and change in winter to summer ratio, respectively.

287 We mentioned at the beginning that this model applies to each monitored site. However, we have multiple  
 288 sites, and counts are often missing for some seasons or even full years. To facilitate the estimation of  
 289 abundance with missing data, we borrow information from sites with counts, by defining a hierarchical  
 290 structure such that:

$$\begin{aligned}\zeta_{tj} &\sim N(0, \sigma_{\zeta t}^2) \\ \epsilon_{tj} &\sim N(0, \sigma_{et}^2)\end{aligned}$$

291 Therefore, random changes at any site and year come from a common distribution of changes across all  
 292 sites for that year. We thus ensure that variation is contained within similar values in most sites. These  
 293 distributions are normal with variances  $\sigma_{\zeta t}^2$  and  $\sigma_{et}^2$  for changes in abundance and winter to summer ratio,  
 294 respectively.

295 We fit these models in R (R Core Team, 2022) with the additional functionality provided by JAGS  
 296 (Plummer, 2003) using the package jagsUI (Kellner, 2021). We work on a Bayesian framework, using  
 297 non-informative priors, and running three chains for 10,000 iterations each. Similar to the occupancy  
 298 models, these are the details of the models we are working with at the time of writing, and they are intended  
 299 to give an idea of the type of model we are using. The modular nature of BIRDIE allows us to update these  
 300 models when necessary and the updated modelling details will be published on the BIRDIE website.  
 301

## 302 Data and model diagnostics

303 The pipeline needs to run for a multitude of species, with different ecological requirements and  
 304 geographical distributions. Therefore, finding a model that suits all species is challenging. Not only  
 305 may a model not be a good fit for a particular species, but the algorithms used for fitting the model may fail  
 306 to converge due to characteristics of the data.

307 In a first control stage, we have defined the minimum requirements that the data should meet to enter the  
 308 model-fitting process. Species that have been observed in five or less pentads in a year are considered to  
 309 not have enough data to inform an occupancy model. Similarly, we chose only those CWAC sites where  
 310 the species of interest has been counted at least ten times between 1993 and 2021, to fit state-space models.  
 311 Otherwise, data tend to be too sparse to assess trends in abundance reliably. These thresholds are based on  
 312 our own experiences working with these data, and they are considered to be the minimum requirements  
 313 for models to converge successfully. However, meeting these requirements does not guarantee model  
 314 convergence or a good fit. To keep track of potential issues arising during model fitting, and to improve the  
 315 algorithms of the pipeline, each time the pipeline runs it generates several reports that are later examined.

316 To decide whether any occupancy or state-space model converges, we calculate the Gelman-Rubin  
 317 (Rhat) diagnostic (Gelman et al., 2014) for each estimated parameter. These diagnostics are then tabulated  
 318 and stored for future revision. Any Rhat value above 1.1 or below 0.9 is considered to represent lack of  
 319 convergence. Distinctive characteristics of the models with convergence issues are explored and addressed  
 320 on a case by case basis, after the pipeline has finished running.

325 the main environmental gradients within our geographical range, without paying too much attention to  
326 multi-collinearity and overfitting. We are therefore cautious about making causal inference or predictions  
327 out of the range of the data, and so should be other users.

## SYSTEMS AND TECHNOLOGY

328 This section describes the servers, database and dashboard technology. We present the process whereby  
329 the data is processed along the pipeline and transformed into indicators that are presented to the final user.  
330 There are multiple formats in which the information is presented to be able to accommodate the needs of  
331 downstream systems. BIRDIE's data, code and outputs are stored and run on three main systems (Figure  
332 5): the Africa Bird Data servers, and the two BIRDIE servers (servers A and B).

333 The Africa Bird Data servers are hosted at the FitzPatrick Institute for African Ornithology (FIAO),  
334 University of Cape Town, and contain the CWAC and ABAP databases. They also serve these data through  
335 an Application Programming Interface (API).

336 BIRDIE's server A is the access point of the final user to the information generated by BIRDIE. This  
337 information is stored in a data mart. At its core, the data mart is a MySQL database (version 8.0.27), a  
338 widely used, open source, relational database management system. Its main objective is to store BIRDIE's  
339 outputs in a way that provides easy and flexible access to the final user. At the same time, the structure of  
340 the database ensures that the inputs and outputs conform to a given standard, and creates security back-ups  
341 for the stored data. The main mechanism BIRDIE uses to present data to the user is through OpenAPI  
342 web services (OpenAPI Specification, Version 3.1.0), which was designed to provide a standard interface  
343 for documenting and exposing APIs. The public web services offered by the OpenAPI give users the  
344 flexibility to access and download data from the database in multiple ways, without being constrained  
345 by specific functionality of a web application. This technology facilitates the integration of BIRDIE's  
346 outputs into other workflows. For the user that is interested in readily accessing the information through  
347 a dashboard, we have deployed a web application, written in HTML5, CSS, and the most common  
348 and popular JavaScript libraries, including OpenLayers (<https://openlayers.org/>) and Plotly  
349 (<https://plotly.com/>). Among other elements (see section 6), the web application features a  
350 map viewer, based on mviewer (<https://mviewer.netlify.app/en/>), a free and open-source  
351 cartographic application, that has an easy-to-use and intuitive interface.

352 If we thought of server A as being the face of the pipeline, server B would be the brain. All the  
353 functionality in this server revolves around statistical modelling. This server connects with the Africa Bird  
354 Data servers to obtain CWAC and ABAP data, and with external systems, such as Google Earth Engine  
355 or SANBI servers to obtain environmental information. It then runs the main analytical modules of the  
356 pipeline, where occupancy and state-space models are fitted. The analytical workflows are supported by an  
357 Intel Xeon Dual 8 core, with 64 GB RAM and an 8 TB hard drive. The model outputs are made available  
358 to server A, where they are incorporated into the data mart, used to compute derived high-level indicators  
359 by aggregation (see section 4), and prepared to be presented to the final users.

## WEB APPLICATION

360 BIRDIE's web application offers four main menus that provide access to the pipeline outputs in different  
361 ways (see Figure 6):

362 1. An exploration map. Through this menu the user is free to explore the different indicators BIRDIE  
363 computes using a map. This spatial framework can be configured to display information layers, such

In addition to convergence, we assess goodness of fit using posterior predictive checks (Doser et al., 2022; Gelman et al., 2014). This procedure compares some quantity of interest calculated using pseudo-data simulated from the model posterior distribution, with that same quantity calculated from the observed data. In a well-fitting model we would expect real and synthetic data to produce similar values. For occupancy models, we produce simulated detection/non-detection data for each site, species and year and compute the expected number of detections out of as many visits as there were in the data. We compare the results of the simulations with the observed number of detections recorded in the data using Chi-square tests (Doser et al., 2022). For state-space models we follow a similar procedure, but instead of simulating detection/non-detection data for one year, we simulate count data for summer and winter, and aggregate these in a single annual count. Results from the goodness of fit Chi-square tests are also tabulated and stored for revision. Significant deviations detected with these tests are addressed for each case individually.

Due to the computational burden of the pipeline, it is not possible to run multiple models for each species, site and year, to perform model selection. Therefore, model selection is performed on a sample of species, selected to have representation of common and scarce taxa, but that are otherwise selected arbitrarily. Our general approach has been to include a rich set of variables that we believe can explain the main environmental gradients within our geographical range, without paying too much attention to multi-collinearity and overfitting. We are therefore cautious about making causal inference or predictions outside of the range of the data, and so should be other users.

## SYSTEMS AND TECHNOLOGY

In this section we describe the technology that underpins the flow of data along the pipeline until it is transformed into indicators that are presented to the BIRDIE user. BIRDIE's data, code and outputs are stored and run on three main systems (Figure 4): the Africa Bird Data servers, and the two BIRDIE servers (servers A and B).

The Africa Bird Data servers are hosted at the FitzPatrick Institute for African Ornithology, University of Cape Town, and contain the CWAC and ABAP databases. They also serve these data through an Application Programming Interface (API).

BIRDIE's server A is the access point of the final user to the information generated by the pipeline. This information is stored in a data mart, which in essence, is a MySQL database (version 8.0.27), a widely used, open source, relational database management system. Its main objective is to store the outputs of BIRDIE's data analyses and provide easy and flexible access to the final user. At the same time, the structure of the database ensures that inputs and outputs conform to a given standard, and creates security back-ups for the stored data. The main mechanism BIRDIE uses to present data to the user is through OpenAPI web services (OpenAPI Specification, Version 3.1.0), which was designed to provide a standard interface for documenting and exposing APIs. The public web services offered by the OpenAPI give users the flexibility to access and download data from the database without being constrained by the specific functionality of a web application. This technology facilitates the integration of BIRDIE's outputs into other workflows. However, for the user that is interested in readily accessing the information through a dashboard, we have deployed a web application, written in HTML5, CSS, and the most common and popular JavaScript libraries, including OpenLayers (<https://openlayers.org/>) and Plotly (<https://plotly.com/>). Among other elements (see section 6), the web application features a map viewer, based on mviewer (<https://mviewer.netlify.app/en/>), a free and open-source cartographic application, that has an easy-to-use and intuitive interface.

364 as occupancy probabilities for ABAP pentads or waterbird abundance at CWAC sites. Users can also  
365 zoom in and out to find the scale that best fits their needs.

366 2. Site and species summaries, are detailed reports elaborated for users specifically interested in sites or  
367 species, rather than in general exploration. These summaries can also be accessed indirectly through  
368 the exploration map.

369 3. Reporting tools. We mentioned in section 2 that BIRDIE was developed to support reports for  
370 national and international conservation programmes. In this menu, users interested in elaborating, or  
371 accessing the information underpinning these reports, will find this information conveniently packed  
372 in programme-specific summaries.

373 4. Web services. Through this menu users can access BIRDIE's API and retrieve its outputs in the most  
374 flexible way. It is through BIRDIE's API that all maps and plots in the web application are produced.  
375 By accessing this functionality directly, users can download the data themselves and incorporate them  
376 into their own workflows.

## DISCUSSION

377 Data on biodiversity and related environmental drivers are collected at increasingly faster rates. Although  
378 these data can be accessed to support decisions at various levels, it can be difficult for decision makers to  
379 extract relevant information in a timely fashion (MacFadyen et al., 2022; P. J. Stephenson et al., 2017).  
380 Here, we introduce BIRDIE, the South African Biodiversity Data Pipeline for Wetlands and Waterbirds; a  
381 data pipeline that aims to provide the information needed for making evidence-based decisions on wetlands  
382 and waterbirds in southern Africa. Target users of BIRDIE include government and public entities that  
383 need to report on the status of wetlands and waterbirds, as well as site managers, and the general public  
384 (e.g., birdwatchers). Apart from data availability and accessibility, obstacles for using biodiversity data  
385 in decision-making include (MacFadyen et al., 2022; P. J. Stephenson et al., 2017): lack of analysis and  
386 interpretation, lack of technical accessibility with excessive use of jargon, and timely use of data.

387 All data used by BIRDIE are freely available, so one of the core objectives of BIRDIE is to facilitate  
388 information uptake by statistically analysing these data and filtering out observational artefacts introduced  
389 during data collection. Uneven sampling efforts, imperfect detection and missing data are all examples of  
390 how data collection methods can affect data, and if undealt with, mislead decision making. Furthermore,  
391 statistical models also provide measures of uncertainty in their estimates, which must be clearly  
392 communicated to the stakeholders. With all their benefits, these statistical analyses require technical  
393 knowledge and are time-consuming. Therefore, having their outputs pre-computed and readily available  
394 could dramatically increase the impact of the data.

395 BIRDIE relies heavily on citizen science projects such as ABAP and CWAC, which poses clear challenges  
396 in terms of uneven efforts and imperfect detection, but also adds the advantage of having the support of  
397 a large community of observers that provides a continuous and steady flow of data. These data inputs  
398 allow us to run the pipeline periodically to keep the indicators updated and timely. Although we would  
399 like to update our indicators more often, for now we only update once per year due to the computational  
400 requirements of the pipeline, and certain characteristics of the data (e.g., CWAC counts are conducted only  
401 twice a year).

402 Our web application was designed to present indicators that are easily accessed, visualised and interpreted,  
403 avoiding unnecessary jargon. At the same time, and for the interested user, we have allocated some space  
404 for clearly explaining the analytical routines used in all the analyses in dedicated sections. In BIRDIE,  
405 we try to follow the Findable, Accessible, Interoperable, Reusable (FAIR) principles (Wilkinson et al.,

362 If we thought of server A as being the face of the pipeline, server B would be the brain. All the  
363 functionality in this server revolves around statistical modelling. This server connects with the Africa Bird  
364 Data servers to obtain CWAC and ABAP data, and with other external systems, such as Google Earth  
365 Engine or SANBI servers to obtain environmental information. It then runs the main analytical modules  
366 of the pipeline, where occupancy and state-space models are fitted. At the time of writing, the analytical  
367 workflows were supported by an Intel Xeon Dual 8 core, with 64 GB RAM and an 8 TB hard drive. The  
368 model outputs are made available to server A, where they are incorporated into the data mart, used to  
369 compute derived high-level indicators by aggregation (see section 4), and prepared to be presented to the  
370 final users.

371 In terms of code structure, the BIRDIE data pipeline consists of several fundamental building blocks  
372 or modules. The first module, which we call the data source layer (Figure 4) hosts and curates the raw  
373 data. The second module, the analysis layer, analyses the data and estimates the fundamental quantities  
374 of interest, like abundance and occurrence of each species at each wetland or pentad. The third module  
375 consists of the data mart where the outputs of the analyses are stored and indicators are aggregated or  
376 disaggregated to multiple scales. The final module serves the information to the user via APIs, web services  
377 and a web application. The modular structure of BIRDIE enables us to maintain and update individual  
378 parts independently. For example, we could replace the current statistical routines with more efficient ones  
379 without changing the other parts of the pipeline. Or we could add new indicators to the data mart layer  
380 without needing to change the statistical routines that produce the underlying components.

## WEB APPLICATION

381 To cater for different user needs, BIRDIE's web application offers four main menus that provide access to  
382 the pipeline outputs in different ways (see Figure 6):

- 383 1. An exploration map. Through this menu the user can explore the different indicators BIRDIE computes  
384 on a map. This spatial framework can be configured to display information layers, such as occupancy  
385 probabilities for ABAP pentads or waterbird abundance at CWAC sites. Users can also zoom in and  
386 out to find the scale that best fits their needs. In addition to this, there are also environmental layers  
387 that can be overlaid to provide context and generate hypotheses on what might be driving the observed  
388 indicators.
- 389 2. Site and species summaries, are detailed reports elaborated for users focused on some sites or species  
390 in particular, rather than in general exploration. At the moment, site summaries are only available for  
391 those sites that have sufficient CWAC data to be included in BIRDIE's data analysis step. These reports  
392 contain a description of the site/species, links to other resources of interest (e.g., to criteria motivating  
393 declaration of Ramsar site or IUCN conservation status) and summaries prepared from BIRDIE's  
394 indicators. These reports can be exported as a document, and BIRDIE's data used for generating the  
395 reports can be accessed through the data mart and downloaded in common formats such as .json or  
396 .csv.
- 397 3. Reporting tools. We mentioned in section 2 that BIRDIE was developed to support reports for  
398 national and international conservation programmes. In this menu, users interested in elaborating, or  
399 accessing the information underpinning these reports, will find this information conveniently packed  
400 in programme-specific summaries. Similar to site and species summaries, reports for conservation  
401 programmes can also be printed, and the data used to compute statistics and create plots can be  
402 downloaded.



403 4. Web services. Through this menu users can access BIRDIE's API and retrieve its outputs in the most  
404 flexible way. It is through BIRDIE's API that all maps and plots in the web application are produced.  
405 By accessing this functionality directly, users can download the data themselves and incorporate them  
406 into their own workflows.

## DISCUSSION

407 Data on biodiversity and related environmental drivers are collected at increasingly faster rates. Although  
408 these data can be accessed to support decisions at various levels, it can be difficult for decision makers to  
409 extract relevant information in a timely fashion (MacFadyen et al., 2022; Stephenson et al., 2017). Apart  
410 from data availability and accessibility, obstacles for using biodiversity data in decision-making include  
411 (MacFadyen et al., 2022; Stephenson et al., 2017): lack of analysis and interpretation, lack of technical  
412 accessibility with excessive use of jargon, and timely use of data. Here, we introduce BIRDIE, the South  
413 African Biodiversity Data Pipeline for Wetlands and Waterbirds; a data pipeline that aims to provide the  
414 information needed for making evidence-based decisions on wetlands and waterbirds in southern Africa.  
415 Target users of BIRDIE include government and public entities that need to report on the status of wetlands  
416 and waterbirds, as well as site managers, and the general public (e.g., birdwatchers).

417 BIRDIE is the first African full biodiversity data pipeline (from raw data to indicators) that we are  
418 aware of at the time of writing. Although biodiversity data portals are proliferating (Saran et al., 2022),  
419 examples of fully operational workflows for computing and displaying biodiversity indicators are still  
420 scarce (but see Brlik et al. 2021, Boyd et al. 2022). Compared to other richer countries, long-term datasets  
421 from biodiversity monitoring programmes are still scarce in many African countries (Proenca et al 2017,  
422 Stephenson et al. 2017). In South Africa we are lucky to have two good bird monitoring programmes that  
423 provide data on waterbirds. However, even these well established programmes can be hampered by lack of  
424 funds and qualified personnel in remote locations, as we can see by the decreasing coverage of the CWAC  
425 project in the last decade (Figure 2). Critical data on the location, structure and dynamics of freshwater  
426 ecosystems are still scarce and highly local. Thus, BIRDIE relies heavily on citizen science projects such  
427 as ABAP and CWAC, which poses clear challenges in terms of uneven efforts and imperfect detection, but  
428 also adds the advantage of having the support of a large community of observers that provides a continuous  
429 and steady flow of data. These data inputs allow us to run the pipeline periodically to keep the indicators  
430 updated and timely. Although we would like to update our indicators more often, at the time of writing we  
431 only update once per year due to the computational requirements of the pipeline, and certain characteristics  
432 of the data (e.g., CWAC counts are conducted only twice a year).

433 All data used by BIRDIE are freely available, so one of the main contributions of BIRDIE is to facilitate  
434 information uptake by statistically analysing these data and filtering out observational artefacts introduced  
435 during data collection. Uneven sampling efforts, imperfect detection and missing data are all examples of  
436 how data collection methods can affect data (Yoccoz et al, 2001), and if undelt with, mislead decision  
437 making. Furthermore, statistical models also provide measures of uncertainty in their estimates, which  
438 must be clearly communicated to the stakeholders (Kissling et al. 2018). With all their benefits, these  
439 statistical analyses require technical knowledge and are time-consuming. Therefore, having their outputs  
440 pre-computed and readily available could dramatically increase the impact of the data. In this context, one  
441 of our main challenges was running models automatically and periodically for multiple species, which  
442 requires pre-defining and using similar models for all species. Therefore we faced a trade-off between  
443 having accurate models for individual species and having a pipeline that works reasonably well for all  
444 species in general. Users should keep in mind this compromise, and think of BIRDIE's outputs as useful  
445 approximations rather than accurate estimates. We recommend designing bespoke models for those species

406 2016), making all processes reproducible and transparent. All the code used by the pipeline is public, freely  
407 available and based on open-source software.

408 Whenever possible we allow the user to display multiple pieces of information in the same screen,  
409 including some environmental data that we consider relevant, to facilitate information contextualization.  
410 For certain specific users, namely those interested in the main national and international reporting channels  
411 (e.g., Ramsar, AEWA, NBA), we prepare dedicated reports that combine the information that suits their  
412 needs. The contents of these reports have been defined during stakeholder engagement, and are intended to  
413 be dynamically adapted to new necessities.

414 The BIRDIE data pipeline consists of several fundamental building blocks. The first block, which we call  
415 the data source layer (Figure 5) hosts and curates the raw data. Vetting and data cleaning also happens in  
416 this block. The second block, the analysis layer, analyses the data and estimates the fundamental quantities  
417 of interest, like abundance and occurrence of each species at each wetland. The third block consists of  
418 a data mart where the outputs of the analysis are stored and indicators are aggregated or disaggregated  
419 to multiple scales. The final block serves the information to the user via a web application. The modular  
420 structure of BIRDIE enables us to maintain and update individual parts independently. For example, we  
421 could replace the current statistical routines with more efficient ones without changing the other parts of  
422 the pipeline. Or we could add new indicators to the data mart layer without needing to change the statistical  
423 routines that produce the underlying components.

424 Running models automatically and periodically for multiple species requires pre-defining models and  
425 using similar models for all species. Therefore there is a trade-off between having accurate models that  
426 fit individual species well and having a pipeline that works well for all species in general. Users should  
427 keep in mind this compromise, and think of BIRDIE's outputs as useful approximations rather than precise  
428 estimates. We recommend designing bespoke models for those species for which precision is required.

429 In the next phase, we intend to develop more profound links between waterbird population indicators and  
430 wetlands. Waterbirds are often regarded as good indicators of wetland biodiversity and condition. However,  
431 this assumption is rarely proven empirically, and it is apparent that it needs careful consideration on a  
432 case by case basis (Amat and Green, 2010). With advances in the accessibility to biodiversity data, we are  
433 now in a better position to investigate whether these claims hold, and if so, under which conditions. Data  
434 portals such as GBIF.org and in South Africa, the Freshwater Biodiversity Information System (FBIS),  
435 and SANBI's biodiversity data portal. SANBI has also been developing the wetland condition map that  
436 provides information about the ecological status of wetlands, looking at elements such as water quality and  
437 vegetation. These additional sources of information could help us understand how waterbird occurrence,  
438 abundance and diversity relates to the general ecological condition of the hosting wetlands.

439 We will also extend BIRDIE's functionality to cover other African countries with similar available data,  
440 such as Kenya and Nigeria that also use the ABAP protocol. There is also a wealth of information that  
441 BIRDIE has not yet used, such as eBird or iNaturalist, that could improve the outputs of the pipeline.  
442 However, integrating data sources with different sampling designs, coverages and biases is not trivial. Data  
443 integration is a very active topic in the field of statistical ecology. Approaches to combining data range  
444 from pooling multiple data sources together disregarding their different assumptions and biases, to much  
445 more accurate integrated models in which characteristics of each data source are explicitly accounted  
446 for (Fletcher et al., 2019). Although at the expense of increased model complexity, with the application  
447 of newly-developed statistical methods for data integration, we can now explore how different species  
448 interrelate, and inform more effective and efficient conservation actions.

446 for which accuracy is required. Similarly, rare species are likely to appear too sparsely in datasets designed  
447 for monitoring common species for models to work well (Bellingham et al., 2020). For these species, we  
448 should design monitoring protocols and models that are tailored for them. Setting up feedback channels  
449 whereby users can suggest model improvements (e.g., relevant covariates) for certain species is a possible  
450 avenue for development in BIRDIE. However, in this first phase, the idea is to create a baseline pipeline in  
451 which the model structure is similar to all sites and species.

452 In addition, model structure was not designed for making causal inference and therefore confounders  
453 could mislead the user to believe that certain variables are driving emerging patterns, when there is only a  
454 correlation (Stewart et al., 2022). To avoid misinterpretation by the casual user, we favoured displaying  
455 environmental layers that can overlay with model state estimations, rather than presenting marginal  
456 covariate effects estimated by the model. In future versions of BIRDIE, we might consider presenting this  
457 type of information in specific sections with extensive explanations on how to interpret it. The current  
458 version of BIRDIE has a portal that presents indicators that are easily accessed, visualised and interpreted,  
459 avoiding unnecessary jargon. At the same time, and for the interested user, we have allocated some space  
460 for clearly explaining the analytical routines used in all the analyses in dedicated sections. In BIRDIE, we  
461 followed the Findable, Accessible, Interoperable, Reusable (FAIR) principles (Wilkinson et al., 2016),  
462 making all processes reproducible and transparent. All the code used by the pipeline is public, freely  
463 available (<https://github.com/AfricaBirdData>) and based on open-source software.

464 In BIRDIE we envision several avenues for further development. Integration of multiple EBVs into  
465 a common assessment has important advantages for understanding drivers of change and designing  
466 conservation interventions (Bellingham et al. 2020). In the next phase, we intend to develop more profound  
467 links between waterbird population indicators and wetlands. Waterbirds are often regarded as good  
468 indicators of wetland biodiversity and condition. However, this assumption is rarely proven empirically,  
469 and it is apparent that it needs careful consideration on a case by case basis (Amat and Green, 2010). With  
470 advances in the accessibility to biodiversity data, we are now in a better position to investigate whether  
471 these claims hold, and if so, under which conditions. Data portals such as GBIF.org and in South Africa, the  
472 Freshwater Biodiversity Information System (FBIS), and SANBI's biodiversity data portal, could help us  
473 understand how waterbird occurrence, abundance and diversity relates to the general ecological condition  
474 of the hosting wetlands. However, we are aware that the integration of opportunistic data with different  
475 sampling schemes and scales poses additional challenges that we will need to carefully address (Kissling et  
476 al. 2018, Boyd et al. 2022)

477 We will also extend BIRDIE's functionality to cover other African countries with similar available data,  
478 such as Kenya and Nigeria that also use the ABAP protocol. There is also a wealth of information that  
479 BIRDIE has not yet used, such as eBird or iNaturalist, that could improve the outputs of the pipeline. While  
480 integrating data sources with different sampling designs, coverages and biases is not trivial, the modular  
481 design of BIRDIE allows us to update the modelling step as new statistical methods are being developed.  
482 Data integration is a very active topic in the field of statistical ecology (Isaac et al., 2020). Approaches to  
483 combining data range from pooling multiple data sources together disregarding their different assumptions  
484 and biases, to much more accurate integrated models in which characteristics of each data source are  
485 explicitly accounted for (Fletcher et al., 2019). Although at the expense of increased model complexity,  
486 with the application of newly-developed statistical methods for data integration, we can now explore how  
487 different species interrelate, and inform more effective and efficient conservation actions.

488 We wish BIRDIE can contribute to closing the existing gap between data providers and decision makers,  
489 facilitating effective conservation action. We also hope it will provide a feedback channel to SABAP,

## CONFLICT OF INTEREST

449 The authors declare that the research was conducted in the absence of any commercial or financial  
450 relationships that could be construed as a potential conflict of interest.

## AUTHOR CONTRIBUTIONS

451 All authors contributed to conception and design of the project. NJ is the project director; FC, RA, and  
452 VV developed the analyses of the pipeline; NJ, AS and DH work on reporting and indicators; FS and YS,  
453 designed and implement the data mart and web application; MB manages the citizen science database. FC  
454 lead the writing of the manuscript with contribution from all authors, who also revised, read, and approved  
455 the submitted version.

## FUNDING

456 This project is funded by the JRS Biodiversity Foundation, grant number 60908.

## ACKNOWLEDGMENTS

457 We are really grateful to other members of the BIRDIE team without whom this project would not be  
458 viable: Sediqa Khatieb, Monica Klass, and Carol Poole. We are also grateful for the support of the JRS  
459 Biodiversity Foundation, and to the many interested users that have engaged and shared useful insights  
460 with us. Finally, we would like to recognize the tremendous contribution of all the citizen scientists that  
461 devote their time and effort to collect the valuable data that we use.

## REFERENCES

- 462 Abatzoglou, J.T., Dobrowski, S.Z., Parks, S.A., Hegewisch, K.C., 2018. TerraClimate, a high-resolution  
463 global dataset of monthly climate and climatic water balance from 1958–2015. *Sci. Data* 5, 170191.  
464 <https://doi.org/10.1038/sdata.2017.191>
- 465 Altwegg, R., Nichols, J.D., 2019. Occupancy models for citizen-science data. *Methods Ecol. Evol.* 10,  
466 8–21. <https://doi.org/10.1111/2041-210X.13090>
- 467 Barnard, P., Altwegg, R., Ebrahim, I., Underhill, L.G., 2017. Early warning systems for biodiversity in  
468 southern Africa – How much can citizen science mitigate imperfect data? *Biol. Conserv.* 208, 183–188.  
469 <https://doi.org/10.1016/j.biocon.2016.09.011>
- 470 Brooks, M., Rose, S., Altwegg, R., Lee, A.T., Nel, H., Ottosson, U., Retief, E., Reynolds, C., Ryan, P.G.,  
471 Shema, S., Tende, T., Underhill, L.G., Thomson, R.L., 2022. The African Bird Atlas Project: a description  
472 of the project and BirdMap data-collection protocol. *Ostrich* 1–10. <https://doi.org/10.2989/00306525.2022.2125097>
- 474 Buckland, S.T., Newman, K.B., Thomas, L., Koesters, N.B., 2004. State-space models for the  
475 dynamics of wild animal populations. *Ecol. Model.* 171, 157–175. <https://doi.org/10.1016/j.ecolmodel.2003.08.002>
- 477 CBD, 2022. Convention on Biological Diversity [WWW Document]. Conv. Biol. Divers. URL <https://www.cbd.int/> (accessed 12.22.22).
- 479 Convention on Wetlands, 2021. Global Wetland Outlook: Special Edition 2021. Secretariat of the  
480 Convention on Wetlands, Gland, Switzerland.

490 CWAC, SANBI's Freshwater Biodiversity Programme and other data providers. Not only serving as a  
491 platform to analyse the data collected, but also to investigate coverage deficiencies and potential new  
492 priorities. Finally, we would like to see that BIRDIE exposes the importance of existing monitoring  
493 programmes, and that it helps prioritise new data-driven initiatives to understand and protect freshwater  
494 biodiversity.

## CONFLICT OF INTEREST

495 The authors declare that the research was conducted in the absence of any commercial or financial  
496 relationships that could be construed as a potential conflict of interest.

## AUTHOR CONTRIBUTIONS

497 All authors contributed to conception and design of the project. NJ is the project director; FC, RA, and VV  
498 developed the analyses of the pipeline; NJ, AS and DH worked on reporting and indicators; FS and YS,  
499 designed and implemented the data mart, APIs, web services and web application; MB manages the citizen  
500 science database. FC led the writing of the manuscript with contribution from all authors, who also revised,  
501 read, and approved the submitted version.

## FUNDING

502 This project is funded by the JRS Biodiversity Foundation, grant number 60908.

## ACKNOWLEDGMENTS

503 We are really grateful to other members of the BIRDIE team without whom this project would not be  
504 viable: Sediqa Khatieb, Monica Klass, and Carol Poole. We are also grateful for the support of the JRS  
505 Biodiversity Foundation, and to the many interested users that have engaged and shared useful insights  
506 with us. Finally, we would like to recognize the tremendous contribution of all the citizen scientists that  
507 devote their time and effort to collect the valuable data that we use.

## REFERENCES

- 508 Abatzoglou, J.T., Dobrowski, S.Z., Parks, S.A., Hegewisch, K.C., 2018. TerraClimate, a high-resolution  
509 global dataset of monthly climate and climatic water balance from 1958–2015. Sci. Data 5, 170191.  
510 <https://doi.org/10.1038/sdata.2017.191>
- 511 Altwegg, R., Nichols, J.D., 2019. Occupancy models for citizen-science data. Methods Ecol. Evol. 10,  
512 8–21. <https://doi.org/10.1111/2041-210X.13090>
- 513 Barnard, P., Altwegg, R., Ebrahim, I., Underhill, L.G., 2017. Early warning systems for biodiversity in  
514 southern Africa – How much can citizen science mitigate imperfect data? Biol. Conserv. 208, 183–188.  
515 <https://doi.org/10.1016/j.biocon.2016.09.011>
- 516 Bellingham, Peter J., Sarah J. Richardson, Andrew M. Gormley, Robert B. Allen, Asher Cook, Philippa  
517 N. Crisp, David M. Forsyth, et al. 2020. ‘Implementing Integrated Measurements of Essential Biodiversity  
518 Variables at a National Scale’. Ecological Solutions and Evidence 1 (2). <https://doi.org/10.1002/2688-8319.12025>.
- 520 Boyd, R., August, T., Cooke, R., Logie, M., Mancini, F., Powney, G., Roy, D., Turvey, K., Isaac, N. 2022.  
521 An operational workflow for producing periodic estimates of species occupancy at large scales. Pre-print.  
522 EcoEvoRxiv. <https://doi.org/10.32942/osf.io/2v7jp>

- 481 Dallas, H., Shelton, J., Sutton, T., Tri Cuptura, D., Kajee, M., Job, N., 2021. The Freshwater Biodiversity  
482 Information System (FBIS) – mobilising data for evaluating long-term change in South African rivers. Afr.  
483 J. Aquat. Sci. 1–16. <https://doi.org/10.2989/16085914.2021.1982672>
- 484 Didan, Kamel, 2015. MOD13A2 MODIS/Terra Vegetation Indices 16-Day L3 Global 1km SIN Grid  
485 V006. <https://doi.org/10.5067/MODIS/MOD13A2.006>
- 486 Doser, J.W., Finley, A.O., Kéry, M., Zipkin, E.F., 2022. spOccupancy: An R package for single-  
487 species, multi-species, and integrated spatial occupancy models. Methods Ecol. Evol. 13, 1670–1678.  
488 <https://doi.org/10.1111/2041-210X.13897>
- 489 FIAO, F.I. of A.O., 2022. CWAC: Coordinated Waterbird Counts [WWW Document]. URL <https://cwac.birdmap.africa/index.php> (accessed 12.21.22).
- 490
- 491 Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D., 2014. Bayesian Data Analysis.  
492 CRC Press, Taylor and Francis Group, Boca Raton, FL.
- 493 Gimenez, O., Buckland, S.T., Morgan, B.J.T., Bez, N., Bertrand, S., Choquet, R., Dray, S., Etienne, M.-P.,  
494 Fewster, R., Gosselin, F., Mérigot, B., Monestiez, P., Morales, J.M., Mortier, F., Munoz, F., Ovaskainen,  
495 O., Pavoline, S., Pradel, R., Schurr, F.M., Thomas, L., Thuiller, W., Trenkel, V., de Valpine, P., Rexstad,  
496 E., 2014. Statistical ecology comes of age. Biol. Lett. 10, 20140698. <https://doi.org/10.1098/rsbl.2014.0698>
- 497
- 498 Han, X., Josse, C., Young, B.E., Smyth, R.L., Hamilton, H.H., Bowles-Newark, N., 2017. Monitoring  
499 national conservation progress with indicators derived from global and national datasets. Biol. Conserv.  
500 213, 325–334. <https://doi.org/10.1016/j.biocon.2016.08.023>
- 501 IUCN, 2022. International Union for the Conservation of Nature [WWW Document]. IUCN. URL  
502 <https://www.iucn.org/content/home-page> (accessed 12.22.22).
- 503 Jetz, W., McGeoch, M.A., Guralnick, R., Ferrier, S., Beck, J., Costello, M.J., Fernandez, M., Geller,  
504 G.N., Keil, P., Merow, C., Meyer, C., Muller-Karger, F.E., Pereira, H.M., Regan, E.C., Schmeller, D.S.,  
505 Turak, E., 2019. Essential biodiversity variables for mapping and monitoring species populations. Nat.  
506 Ecol. Evol. 3, 539–551. <https://doi.org/10.1038/s41559-019-0826-1>
- 507 Kellner, K., 2021. jagsUI: A Wrapper Around “rjags” to Streamline “JAGS” Analyses.
- 508 King, R., 2014. Statistical Ecology. Annu. Rev. Stat. Its Appl. 1, 401–426. <https://doi.org/10.1146/annurev-statistics-022513-115633>
- 509
- 510 Kissling, W.D., Ahumada, J.A., Bowser, A., Fernandez, M., Fernández, N., García, E.A., Guralnick, R.P.,  
511 Isaac, N.J.B., Kelling, S., Los, W., McRae, L., Mihoub, J., Obst, M., Santamaria, M., Skidmore, A.K.,  
512 Williams, K.J., Agosti, D., Amariles, D.,
- 513 Arvanitidis, C., Bastin, L., De Leo, F., Egloff, W., Elith, J., Hobern, D., Martin, D., Pereira, H.M., Pesole,  
514 G., Peterseil, J., Saarenmaa, H., Schigel, D., Schmeller, D.S., Segata, N., Turak, E., Uhlir, P.F., Wee,  
515 B., Hardisty, A.R., 2018. Building essential biodiversity variables (EBVs) of species distribution and  
516 abundance at a global scale. Biol. Rev. 93, 600–625. <https://doi.org/10.1111/brv.12359>
- 517 Mace, G.M., Barrett, M., Burgess, N.D., Cornell, S.E., Freeman, R., Grooten, M., Purvis, A., 2018.  
518 Aiming higher to bend the curve of biodiversity loss. Nat. Sustain. 1, 448–451. <https://doi.org/10.1038/s41893-018-0130-0>
- 519

- 523 Brlík, Vojtěch, Eva Šilarová, Jana Škorpilová, Hany Alonso, Marc Anton, Ainars Aunins, Zoltán  
524 Benkő, et al. 2021. ‘Long-Term and Large-Scale Multispecies Dataset Tracking Population Changes  
525 of Common European Breeding Birds’. *Scientific Data* 8 (1): 21. <https://doi.org/10.1038/s41597-021-00804-2>.
- 527 Brooks, M., Rose, S., Altwegg, R., Lee, A.T., Nel, H., Ottosson, U., Retief, E., Reynolds, C., Ryan, P.G.,  
528 Shema, S., Tende, T., Underhill, L.G., Thomson, R.L., 2022. The African Bird Atlas Project: a description  
529 of the project and BirdMap data-collection protocol. *Ostrich* 1–10. <https://doi.org/10.2989/00306525.2022.2125097>
- 531 Buckland, S.T., Newman, K.B., Thomas, L., Koesters, N.B., 2004. State-space models for the  
532 dynamics of wild animal populations. *Ecol. Model.* 171, 157–175. <https://doi.org/10.1016/j.ecolmodel.2003.08.002>
- 534 CBD, 2022. Convention on Biological Diversity [WWW Document]. Conv. Biol. Divers. URL <https://www.cbd.int/> (accessed 12.22.22).
- 536 Convention on Wetlands, 2021. Global Wetland Outlook: Special Edition 2021. Secretariat of the  
537 Convention on Wetlands, Gland, Switzerland.
- 538 Dallas, H., Shelton, J., Sutton, T., Tri Cuptura, D., Kajee, M., Job, N., 2021. The Freshwater Biodiversity  
539 Information System (FBIS) – mobilising data for evaluating long-term change in South African rivers. *Afr.  
540 J. Aquat. Sci.* 1–16. <https://doi.org/10.2989/16085914.2021.1982672>
- 541 Didan, Kamel, 2015. MOD13A2 MODIS/Terra Vegetation Indices 16-Day L3 Global 1km SIN Grid  
542 V006. <https://doi.org/10.5067/MODIS/MOD13A2.006>
- 543 Doser, J.W., Finley, A.O., Kéry, M., Zipkin, E.F., 2022. spOccupancy: An R package for single-  
544 species, multi-species, and integrated spatial occupancy models. *Methods Ecol. Evol.* 13, 1670–1678.  
545 <https://doi.org/10.1111/2041-210X.13897>
- 546 FIAO, F.I. of A.O., 2022. CWAC: Coordinated Waterbird Counts [WWW Document]. URL <https://cwac.birdmap.africa/index.php> (accessed 12.21.22).
- 548 Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D., 2014. Bayesian Data Analysis.  
549 CRC Press, Taylor and Francis Group, Boca Raton, FL.
- 550 Gimenez, O., Buckland, S.T., Morgan, B.J.T., Bez, N., Bertrand, S., Choquet, R., Dray, S., Etienne, M.-P.,  
551 Fewster, R., Gosselin, F., Mérigot, B., Monestiez, P., Morales, J.M., Mortier, F., Munoz, F., Ovaskainen,  
552 O., Pavoline, S., Pradel, R., Schurr, F.M., Thomas, L., Thuiller, W., Trenkel, V., de Valpine, P., Rexstad,  
553 E., 2014. Statistical ecology comes of age. *Biol. Lett.* 10, 20140698. <https://doi.org/10.1098/rsbl.2014.0698>
- 555 Han, X., Josse, C., Young, B.E., Smyth, R.L., Hamilton, H.H., Bowles-Newark, N., 2017. Monitoring  
556 national conservation progress with indicators derived from global and national datasets. *Biol. Conserv.*  
557 213, 325–334. <https://doi.org/10.1016/j.biocon.2016.08.023>
- 558 IUCN, 2022. International Union for the Conservation of Nature [WWW Document]. IUCN. URL  
559 <https://www.iucn.org/content/home-page> (accessed 12.22.22).
- 560 Jetz, W., McGeoch, M.A., Guralnick, R., Ferrier, S., Beck, J., Costello, M.J., Fernandez, M., Geller,  
561 G.N., Keil, P., Merow, C., Meyer, C., Muller-Karger, F.E., Pereira, H.M., Regan, E.C., Schmeller, D.S.,

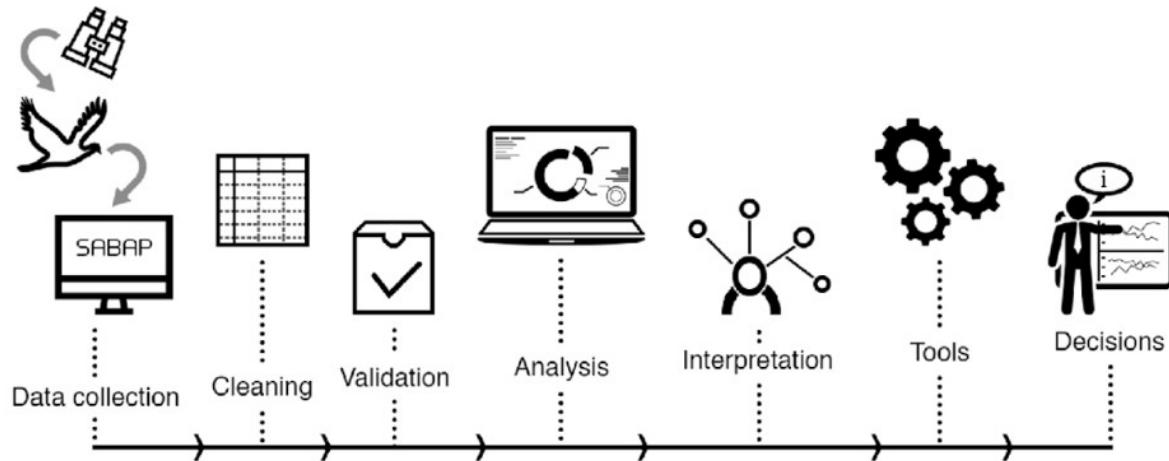
- 520 MacFadyen, S., Allsopp, N., Altwegg, R., Archibald, S., Botha, J., Bradshaw, K., Carruthers, J., De Clerk,  
521 H., de Vos, A., Distiller, G., Foord, S., Freitag-Ronaldson, S., Gibbs, R., Hamer, M., Landi, P., MacFadyen,  
522 D., Manuel, J., Midgley, G., Moncrieff, G., Munch, Z., Mutanga, O., Sershen, Nenguda, R., Ngwenya,  
523 M., Parker, D., Peel, M., Power, J., Pretorius, J., Ramdhani, S., Robertson, M., Rushworth, I., Skowno,  
524 A., Slingsby, J., Turner, A., Visser, V., Van Wageningen, G., Hui, C., 2022. Drowning in data, thirsty for  
525 information and starved for understanding: A biodiversity information hub for cooperative environmental  
526 monitoring in South Africa. *Biol. Conserv.* 274, 109736. <https://doi.org/10.1016/j.biocon.2022.109736>
- 528 MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Andrew Royle, J., Langtimm, C.A., 2002.  
529 Estimating site occupancy rates when detection probabilities are less than one. *Ecology* 83, 2248–2255.  
530 [https://doi.org/10.1890/0012-9658\(2002\)083%5B2248:ESORWD%5D2.0.CO;2](https://doi.org/10.1890/0012-9658(2002)083%5B2248:ESORWD%5D2.0.CO;2)
- 531 Newman, K.B., Buckland, S.T., Morgan, B.J.T., King, R., Borchers, D.L., Cole, D.J., Besbeas,  
532 P., Gimenez, O., Thomas, L., 2014. Modelling Population Dynamics: model formulation, fitting and  
533 assessment using state-space methods. Springer, New York, NY.
- 534 Pekel, J.-F., Cottam, A., Gorelick, N., Belward, A.S., 2016. High-resolution mapping of global  
535 surface water and its long-term changes. *Nature* 540, 418–422. <https://doi.org/10.1038/nature20584>
- 537 Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., Bruford, M.W.,  
538 Brummitt, N., Butchart, S.H.M., Cardoso, A.C., Coops, N.C., Dulloo, E., Faith, D.P., Freyhof, J., Gregory,  
539 R.D., Heip, C., Höft, R., Hurt, G., Jetz, W., Karp, D.S., McGeoch, M.A., Obura, D., Onoda, Y., Pettorelli,  
540 N., Reyers, B., Sayre, R., Scharlemann, J.P.W., Stuart, S.N., Turak, E., Walpole, M., Wegmann, M., 2013.  
541 Essential Biodiversity Variables. *Science* 339, 277–278. <https://doi.org/10.1126/science.1229931>
- 543 Plummer, M., 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling,  
544 in: Proceedings of the 3rd International Workshop on Distributed Statistical Computing.
- 545 Pollock, L.J., O'Connor, L.M.J., Mokany, K., Rosauer, D.F., Talluto, M.V., Thuiller, W., 2020. Protecting  
546 Biodiversity (in All Its Complexity): New Models and Methods. *Trends Ecol. Evol.* 35, 1119–1128.  
547 <https://doi.org/10.1016/j.tree.2020.08.015>
- 548 R Core Team, 2022. R: A Language and Environment for Statistical Computing.
- 549 SANBI, South African National Biodiversity Institute, 2023. Biodiversity Advisor [WWW Document].  
550 URL <http://biodiversityadvisor.sanbi.org/> (accessed 12.21.22).
- 551 SANBI, South African National Biodiversity Institute, in prep. National Wetland Map version 6.
- 552 Skowno, A., Poole, C.J., Raimondo, D.C., Sink, K.J., Van Deventer, H., Van Niekerk, L., Harris, L.R.,  
553 Smith-Adao, L.B., Tolley, K.A., Zengeya, T.A., Foden, W.B., Midgley, G.F., Driver, A., 2019. National  
554 biodiversity assessment 2018: the status of South Africa's ecosystems and biodiversity: synthesis report.  
555 South African National Biodiversity Institute, Department of Environment, Forestry and Fisheries, Pretoria.
- 556 Stephenson, P., Brooks, T.M., Butchart, S.H., Fegraus, E., Geller, G.N., Höft, R., Hutton, J., Kingston,  
557 N., Long, B., McRae, L., 2017. Priorities for big biodiversity data. *Front. Ecol. Environ.* 15, 124–125.  
558 <https://doi.org/10.1002/fee.1473>
- 559 Stephenson, P.J., Bowles-Newark, N., Regan, E., Stanwell-Smith, D., Diagana, M., Höft, R., Abarchi, H.,  
560 Abrahamse, T., Akello, C., Allison, H., Banki, O., Batieno, B., Dieme, S., Domingos, A., Galt, R., Githaiga,

- 562 Turak, E., 2019. Essential biodiversity variables for mapping and monitoring species populations. *Nat.*  
563 *Ecol. Evol.* 3, 539–551. <https://doi.org/10.1038/s41559-019-0826-1>
- 564 Kellner, K., 2021. jagsUI: A Wrapper Around “rjags” to Streamline “JAGS” Analyses.
- 565 King, R., 2014. Statistical Ecology. *Annu. Rev. Stat. Its Appl.* 1, 401–426. <https://doi.org/10.1146/annurev-statistics-022513-115633>
- 567 Kissling, W.D., Ahumada, J.A., Bowser, A., Fernandez, M., Fernández, N., García, E.A., Guralnick,  
568 R.P., Isaac, N.J.B., Kelling, S., Los, W., McRae, L., Mihoub, J., Obst, M., Santamaria, M., Skidmore,  
569 A.K., Williams, K.J., Agosti, D., Amariles, D., Arvanitidis, C., Bastin, L., De Leo, F., Egloff, W.,  
570 Elith, J., Hobern, D., Martin, D., Pereira, H.M., Pesole, G., Peterseil, J., Saarenmaa, H., Schigel, D.,  
571 Schmeller, D.S., Segata, N., Turak, E., Uhlir, P.F., Wee, B., Hardisty, A.R., 2018. Building essential  
572 biodiversity variables (EBVs) of species distribution and abundance at a global scale. *Biol. Rev.* 93,  
573 600–625. <https://doi.org/10.1111/brv.12359>
- 574 Mace, G.M., Barrett, M., Burgess, N.D., Cornell, S.E., Freeman, R., Grootenhuis, M., Purvis, A., 2018.  
575 Aiming higher to bend the curve of biodiversity loss. *Nat. Sustain.* 1, 448–451. <https://doi.org/10.1038/s41893-018-0130-0>
- 577 MacFadyen, S., Allsopp, N., Altweig, R., Archibald, S., Botha, J., Bradshaw, K., Carruthers, J., De Clerk,  
578 H., de Vos, A., Distiller, G., Foord, S., Freitag-Ronaldson, S., Gibbs, R., Hamer, M., Landi, P., MacFadyen,  
579 D., Manuel, J., Midgley, G., Moncrieff, G., Munch, Z., Mutanga, O., Sershen, Nenguda, R., Ngwenya,  
580 M., Parker, D., Peel, M., Power, J., Pretorius, J., Ramdhani, S., Robertson, M., Rushworth, I., Skowno,  
581 A., Slingsby, J., Turner, A., Visser, V., Van Wageningen, G., Hui, C., 2022. Drowning in data, thirsty for  
582 information and starved for understanding: A biodiversity information hub for cooperative environmental  
583 monitoring in South Africa. *Biol. Conserv.* 274, 109736. <https://doi.org/10.1016/j.biocon.2022.109736>
- 585 MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Andrew Royle, J., Langtimm, C.A., 2002.  
586 Estimating site occupancy rates when detection probabilities are less than one. *Ecology* 83, 2248–2255.  
587 [https://doi.org/10.1890/0012-9658\(2002\)083%5B2248:ESORWD%5D2.0.CO;2](https://doi.org/10.1890/0012-9658(2002)083%5B2248:ESORWD%5D2.0.CO;2)
- 588 Newman, K.B., Buckland, S.T., Morgan, B.J.T., King, R., Borchers, D.L., Cole, D.J., Besbeas, P., Gimenez, O., Thomas, L., 2014. Modelling Population Dynamics: model formulation, fitting and  
589 assessment using state-space methods. Springer, New York, NY.
- 591 Pekel, J.-F., Cottam, A., Gorelick, N., Belward, A.S., 2016. High-resolution mapping of global  
592 surface water and its long-term changes. *Nature* 540, 418–422. <https://doi.org/10.1038/nature20584>
- 594 Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H.G., Scholes, R.J., Bruford, M.W.,  
595 Brummitt, N., Butchart, S.H.M., Cardoso, A.C., Coops, N.C., Dulloo, E., Faith, D.P., Freyhof, J., Gregory,  
596 R.D., Heip, C., Höft, R., Hurt, G., Jetz, W., Karp, D.S., McGeoch, M.A., Obura, D., Onoda, Y., Pettorelli,  
597 N., Reyers, B., Sayre, R., Scharlemann, J.P.W., Stuart, S.N., Turak, E., Walpole, M., Wegmann, M., 2013.  
598 Essential Biodiversity Variables. *Science* 339, 277–278. <https://doi.org/10.1126/science.1229931>
- 600 Plummer, M., 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling,  
601 in: Proceedings of the 3rd International Workshop on Distributed Statistical Computing.

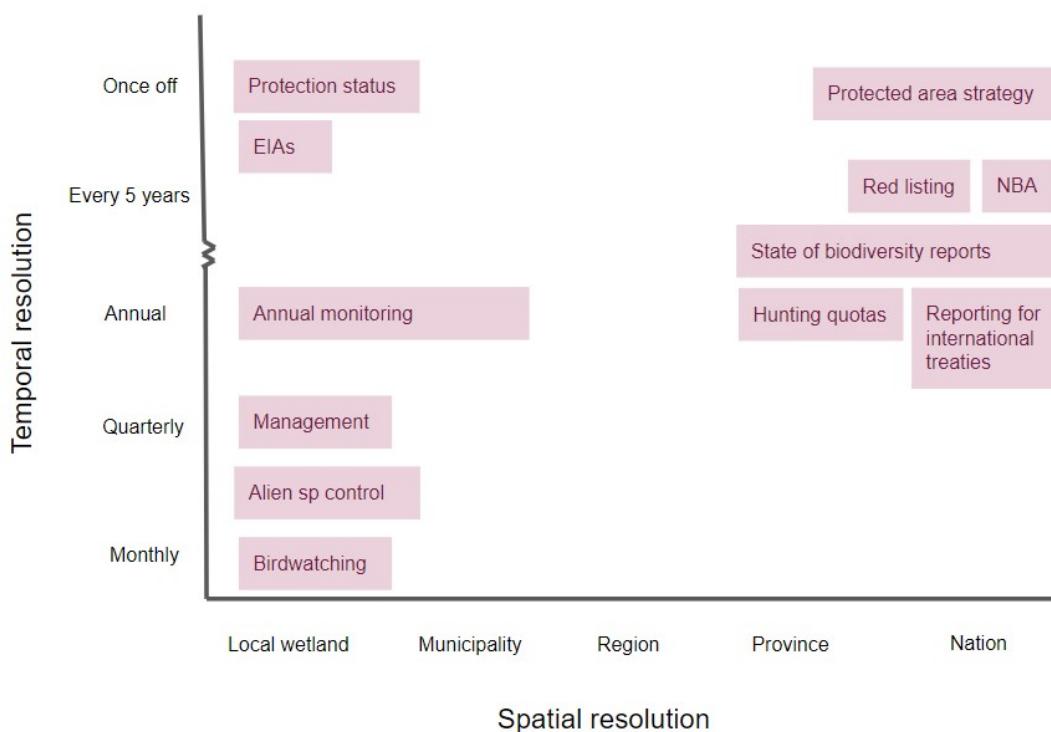
- 561 C.W., Guindo, A.B., Hafashimana, D.L.N., Hirsch, T., Hoborn, D., Kaaya, J., Kaggwa, R., Kalemba, M.M.,  
562 Linjouom, I., Manaka, B., Mbwambo, Z., Musasa, M., Okoree, E., Rwetsiba, A., Siam, A.B., Thiombiano,  
563 A., 2017. Unblocking the flow of biodiversity data for decision-making in Africa. *Biol. Conserv.* 213,  
564 335–340. <https://doi.org/10.1016/j.biocon.2016.09.003>
- 565 Stephenson, P.J., Ntiamoa-Baidu, Y., Simaika, J.P., 2020. The Use of Traditional and Modern Tools for  
566 Monitoring Wetlands Biodiversity in Africa: Challenges and Opportunities. *Front. Environ. Sci.* 8, 1–12.  
567 <https://doi.org/10.3389/fenvs.2020.00061>
- 568 UN, United Nations, 2022. Sustainable Development Goals [WWW Document]. URL <https://sdgs.un.org/> (accessed 12.22.22).
- 570 UNEP, United Nations Environmental Programme, 2022. AEWA: Agreement on the Conservation of  
571 African-Eurasian Migratory Waterbirds [WWW Document]. URL <https://www.unep-aewa.org/>  
572 (accessed 12.22.22).
- 573 Wetzel, F.T., Saarenmaa, H., Regan, E., Martin, C.S., Mergen, P., Smirnova, L., Tuama, É.O., García  
574 Camacho, F.A., Hoffmann, A., Vohland, K., Häuser, C.L., 2015. The roles and contributions of Biodiversity  
575 Observation Networks (BONs) in better tracking progress to 2020 biodiversity targets: a European case  
576 study. *Biodiversity* 16, 137–149. <https://doi.org/10.1080/14888386.2015.1075902>
- 577 White, E.P., Yenni, G.M., Taylor, S.D., Christensen, E.M., Bledsoe, E.K., Simonis, J.L., Ernest, S.K.M.,  
578 2019. Developing an automated iterative near-term forecasting system for an ecological study. *Methods Ecol. Evol.* 10, 332–344. <https://doi.org/10.1111/2041-210X.13104>
- 579 Wilkinson, M.D., Dumontier, M., Aalbersberg, IJ.J., Appleton, G., Axton, M., Baak, A., Blomberg, N.,  
580 Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M.,  
581 Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P.,  
582 Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J.,  
583 Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone,  
584 S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J.,  
585 van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.,  
586 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018.  
587 <https://doi.org/10.1038/sdata.2016.18>
- 588 Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J.C., Sampson, C.C.,  
589 Kanae, S., Bates, P.D., 2017. A high-accuracy map of global terrain elevations: Accurate Global Terrain  
590 Elevation map. *Geophys. Res. Lett.* 44, 5844–5853. <https://doi.org/10.1002/2017GL072874>
- 591 Yenni, G.M., Christensen, E.M., Bledsoe, E.K., Supp, S.R., Diaz, R.M., White, E.P., Ernest, S.K.M.,  
592 2019. Developing a modern data workflow for regularly updated data. *PLOS Biol.* 17, e3000125. <https://doi.org/10.1371/journal.pbio.3000125>
- 593 Yoccoz, N.G., Nichols, J.D., Boulinier, T., 2001. Monitoring of biological diversity in space and time.  
594 Trends Ecol. Evol. 16, 446–453. [https://doi.org/10.1016/S0169-5347\(01\)02205-4](https://doi.org/10.1016/S0169-5347(01)02205-4)

## FIGURES AND CAPTIONS

- 602 Pollock, L.J., O'Connor, L.M.J., Mokany, K., Rosauer, D.F., Talluto, M.V., Thuiller, W., 2020. Protecting  
603 Biodiversity (in All Its Complexity): New Models and Methods. *Trends Ecol. Evol.* 35, 1119–1128.  
604 <https://doi.org/10.1016/j.tree.2020.08.015>
- 605 R Core Team, 2022. R: A Language and Environment for Statistical Computing.
- 606 Saran, Sameer, Sumit Kumar Chaudhary, Priyanka Singh, Amrapali Tiwari, and Vishal Kumar. 2022.  
607 'A Comprehensive Review on Biodiversity Information Portals'. *Biodiversity and Conservation* 31 (5–6):  
608 1445–68. <https://doi.org/10.1007/s10531-022-02420-x>.
- 609 SANBI, South African National Biodiversity Institute, 2023. Biodiversity Advisor [WWW Document].  
610 URL <http://biodiversityadvisor.sanbi.org/> (accessed 12.21.22).
- 611 SANBI, South African National Biodiversity Institute, in prep. National Wetland Map version 6.
- 612 Skowno, A., Poole, C.J., Raimondo, D.C., Sink, K.J., Van Deventer, H., Van Niekerk, L., Harris, L.R.,  
613 Smith-Adao, L.B., Tolley, K.A., Zengeya, T.A., Foden, W.B., Midgley, G.F., Driver, A., 2019. National  
614 biodiversity assessment 2018: the status of South Africa's ecosystems and biodiversity: synthesis report.  
615 South African National Biodiversity Institute, Department of Environment, Forestry and Fisheries, Pretoria.
- 616 Stephenson, P., Brooks, T.M., Butchart, S.H., Fegraus, E., Geller, G.N., Hoft, R., Hutton, J., Kingston,  
617 N., Long, B., McRae, L., 2017. Priorities for big biodiversity data. *Front. Ecol. Environ.* 15, 124–125.  
618 <https://doi.org/10.1002/fee.1473>
- 619 Stephenson, P.J., Bowles-Newark, N., Regan, E., Stanwell-Smith, D., Diagana, M., Höft, R., Abarchi, H.,  
620 Abrahamse, T., Akello, C., Allison, H., Banki, O., Batieno, B., Dieme, S., Domingos, A., Galt, R., Githaiga,  
621 C.W., Guindo, A.B., Hafashimana, D.L.N., Hirsch, T., Hoborn, D., Kaaya, J., Kaggwa, R., Kalemba, M.M.,  
622 Linjouom, I., Manaka, B., Mbawambo, Z., Musasa, M., Okoree, E., Rwetsiba, A., Siam, A.B., Thiombiano,  
623 A., 2017. Unblocking the flow of biodiversity data for decision-making in Africa. *Biol. Conserv.* 213,  
624 335–340. <https://doi.org/10.1016/j.biocon.2016.09.003>
- 625 Stephenson, P.J., Ntiamo-Baidu, Y., Simaika, J.P., 2020. The Use of Traditional and Modern Tools for  
626 Monitoring Wetlands Biodiversity in Africa: Challenges and Opportunities. *Front. Environ. Sci.* 8, 1–12.  
627 <https://doi.org/10.3389/fenvs.2020.00061>
- 628 UN, United Nations, 2022. Sustainable Development Goals [WWW Document]. URL <https://sdgs.un.org/> (accessed 12.22.22).
- 630 UNEP, United Nations Environmental Programme, 2022. AEWA: Agreement on the Conservation of  
631 African-Eurasian Migratory Waterbirds [WWW Document]. URL <https://www.unep-aewa.org/>  
632 (accessed 12.22.22).
- 633 Wetzel, F.T., Saarenmaa, H., Regan, E., Martin, C.S., Mergen, P., Smirnova, L., Tuama, É.O., García  
634 Camacho, F.A., Hoffmann, A., Vohland, K., Häuser, C.L., 2015. The roles and contributions of Biodiversity  
635 Observation Networks (BONs) in better tracking progress to 2020 biodiversity targets: a European case  
636 study. *Biodiversity* 16, 137–149. <https://doi.org/10.1080/14888386.2015.1075902>
- 637 White, E.P., Yenni, G.M., Taylor, S.D., Christensen, E.M., Bledsoe, E.K., Simonis, J.L., Ernest, S.K.M.,  
638 2019. Developing an automated iterative near-term forecasting system for an ecological study. *Methods  
639 Ecol. Evol.* 10, 332–344. <https://doi.org/10.1111/2041-210X.13104>
- 640 Wilkinson, M.D., Dumontier, M., Aalbersberg, IJ.J., Appleton, G., Axton, M., Baak, A., Blomberg, N.,  
641 Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M.,



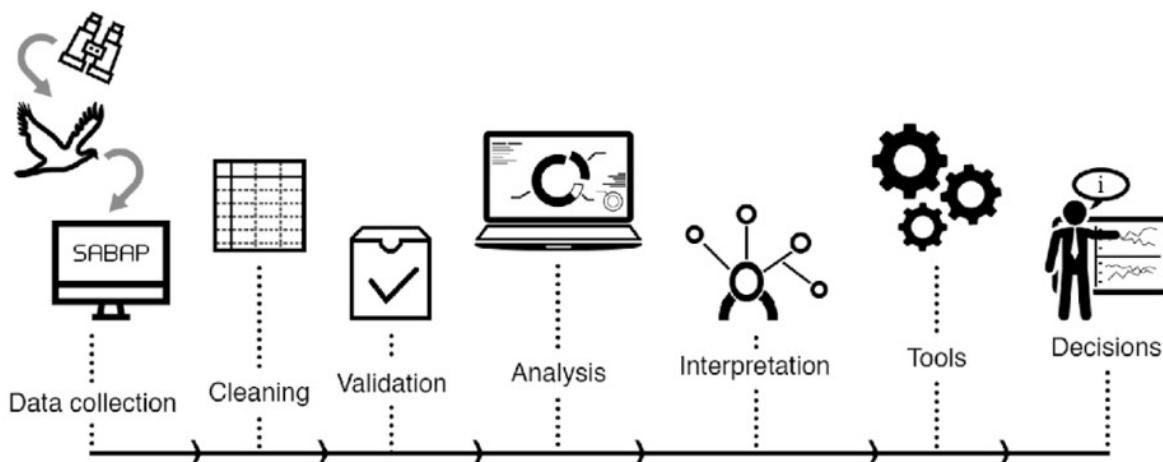
**Figure 1.** Basic workflow of the BIRDIE pipeline covering all steps from data collection, to analysis and presentation of digested, decision-ready indicators.



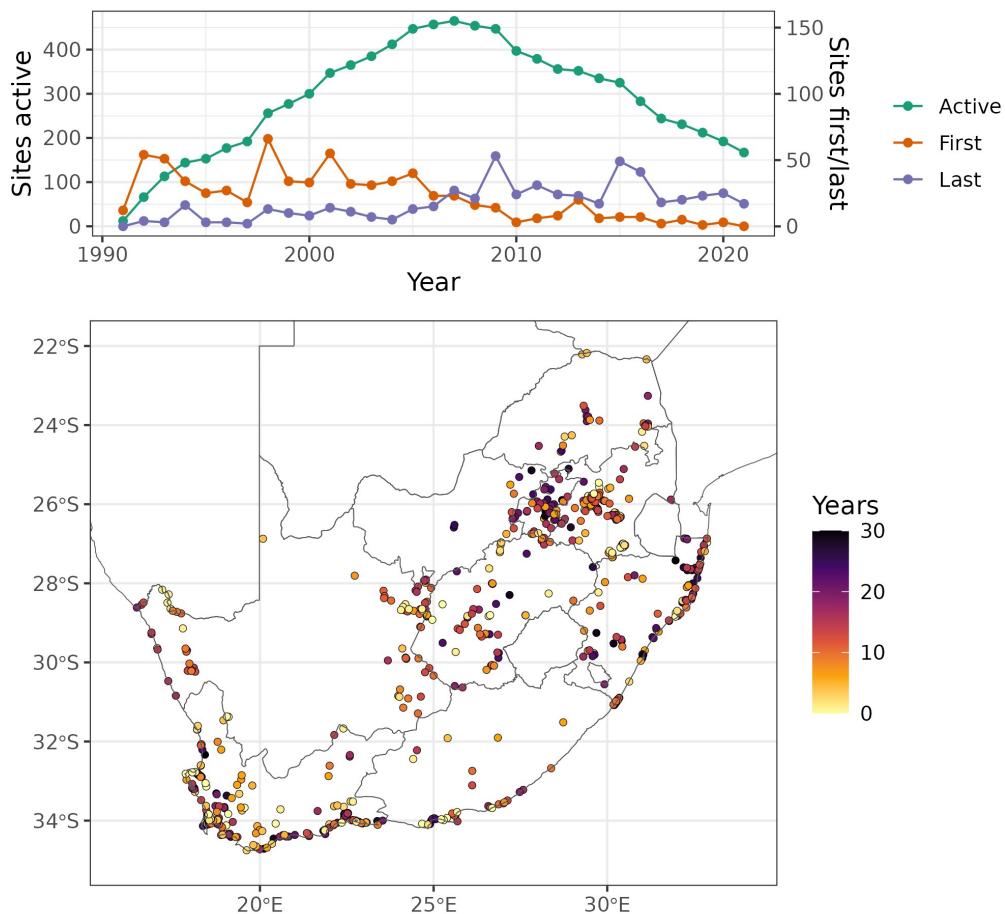
**Figure 2.** Main users targeted by BIRDIE in relation to their spatial and temporal assessment scales. At present, we focus on computing indicators at an annual temporal resolution or coarser. Finer resolutions are typically based on access to raw data.

- 642 Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P.,  
643 Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J.,  
644 Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone,  
645 S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J.,  
646 van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.,  
647 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 160018.  
648 <https://doi.org/10.1038/sdata.2016.18>
- 649 Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J.C., Sampson, C.C.,  
650 Kanae, S., Bates, P.D., 2017. A high-accuracy map of global terrain elevations: Accurate Global Terrain  
651 Elevation map. *Geophys. Res. Lett.* 44, 5844–5853. <https://doi.org/10.1002/2017GL072874>
- 652 Yenni, G.M., Christensen, E.M., Bledsoe, E.K., Supp, S.R., Diaz, R.M., White, E.P., Ernest, S.K.M.,  
653 2019. Developing a modern data workflow for regularly updated data. *PLOS Biol.* 17, e3000125. <https://doi.org/10.1371/journal.pbio.3000125>
- 655 Yoccoz, N.G., Nichols, J.D., Boulinier, T., 2001. Monitoring of biological diversity in space and time.  
656 Trends Ecol. Evol. 16, 446–453. [https://doi.org/10.1016/S0169-5347\(01\)02205-4](https://doi.org/10.1016/S0169-5347(01)02205-4)

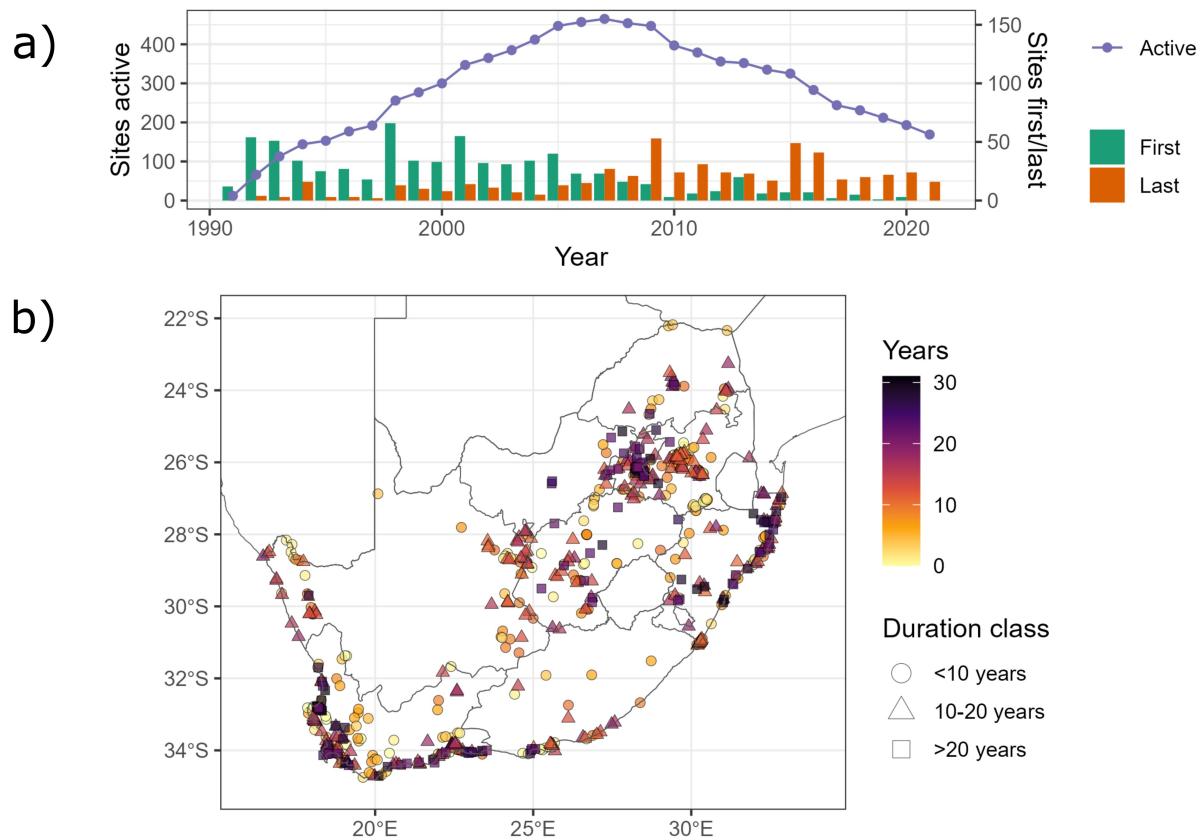
## FIGURES AND CAPTIONS



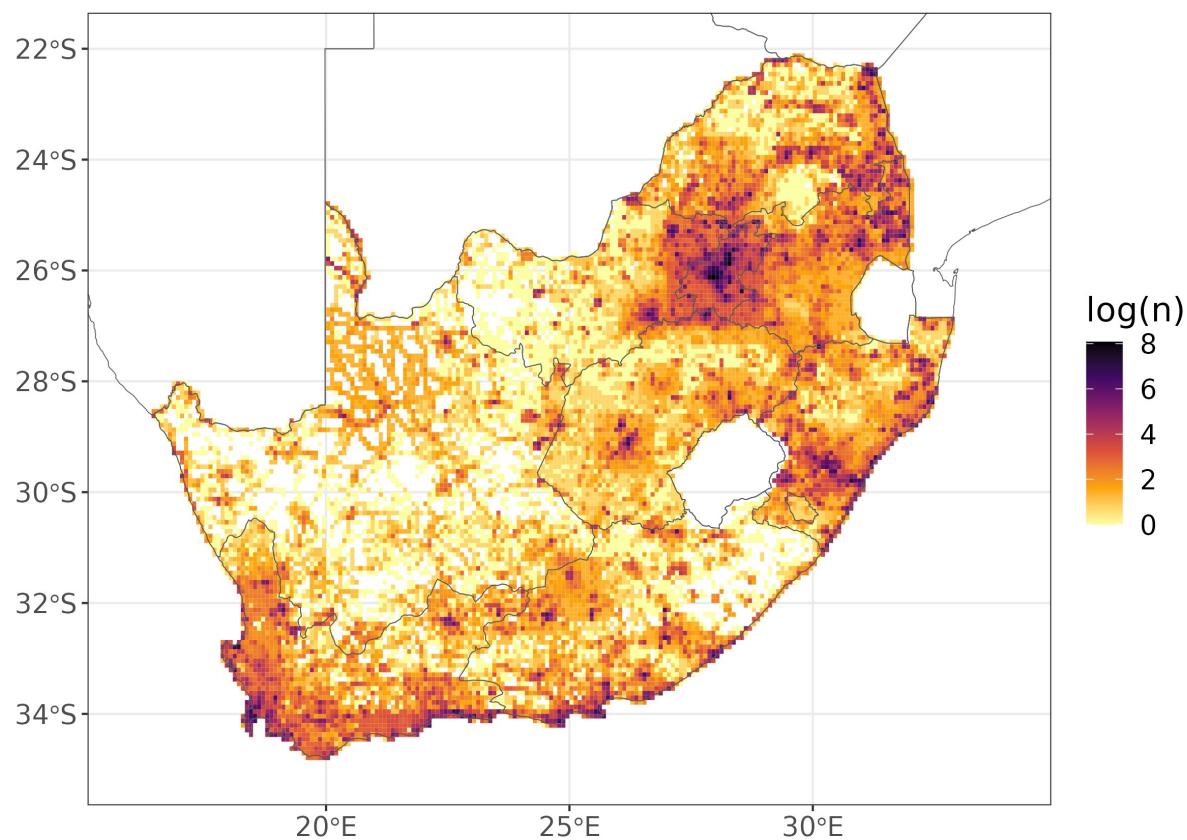
**Figure 1.** Basic workflow of the BIRDIE pipeline covering all steps from data collection, to analysis and presentation of digested, decision-ready indicators. Note that this is not a detailed sequence of all steps data go through, but rather a simplified view of the main processes.



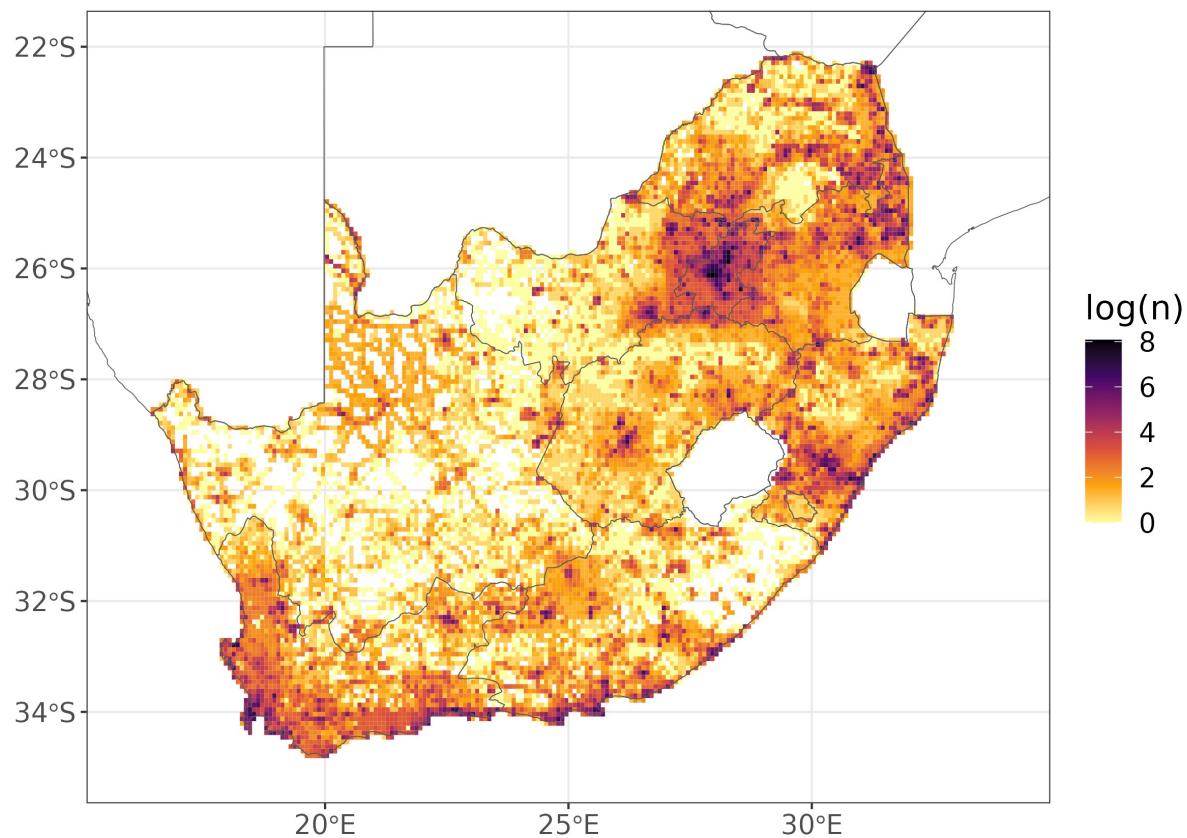
**Figure 3.** The graph shows, the number of CWAC sites active (green), firstly counted (red) and last counted (purple), per year, between 1991 and 2021. Note that some of the sites that were last counted before 2021, might be counted again in the future. In the map, the spatial location of CWAC sites in South Africa. The colour gradient represent the duration of the period the site was counted for.



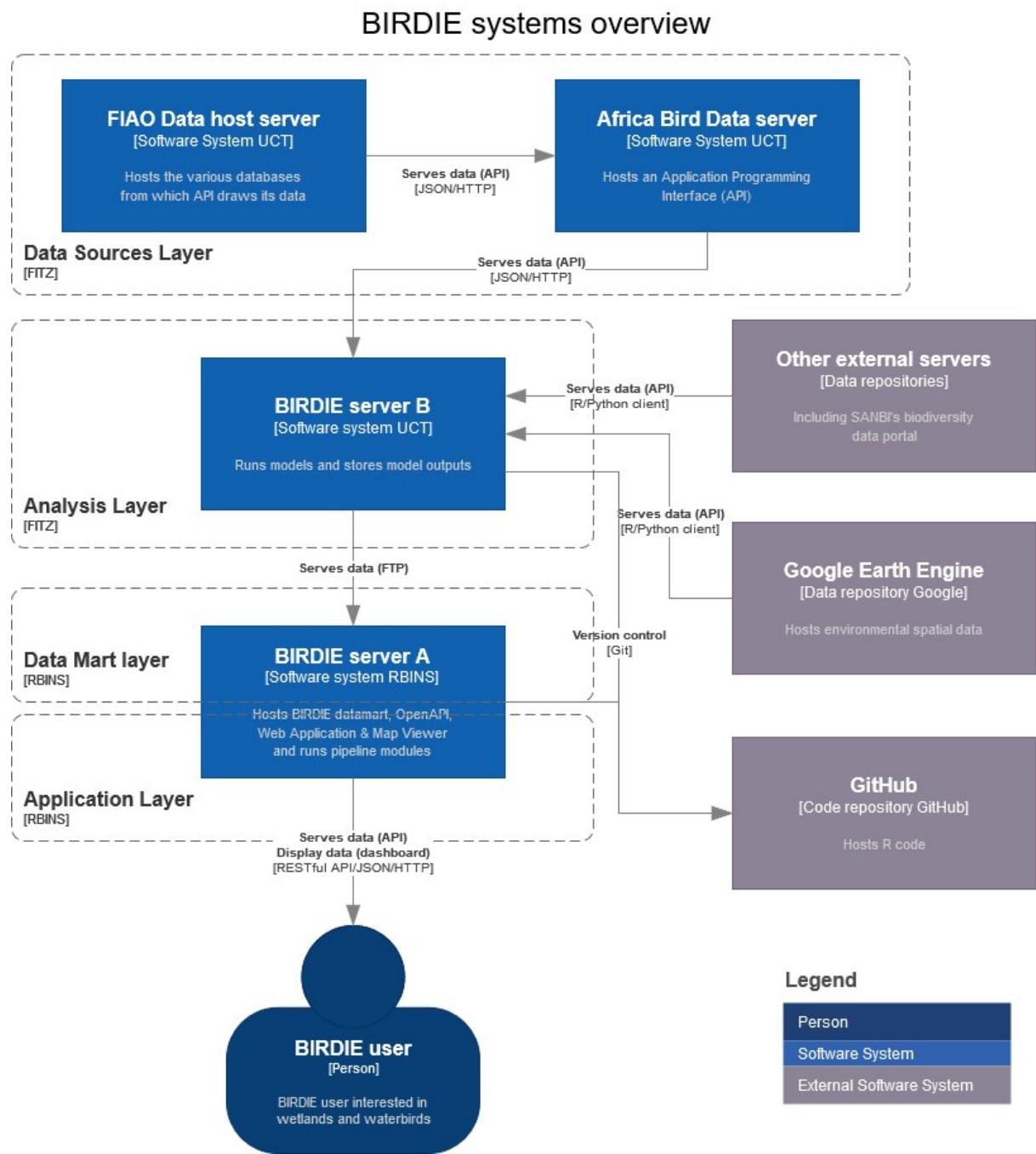
**Figure 2.** The graph a) shows, the number of CWAC sites active (purple dots and line), number of sites firstly counted each year (green bats) and number of sites last counted each year (orange bars), between 1991 and 2021. Note that some of the sites that were last counted before 2021, might be counted again in the future, so orange bars do not represent sites removed from the programme. In map b), we show the spatial distribution of CWAC sites in South Africa. The colour gradient represent the duration of the period the site was counted for. To aid visualisation, we show different shapes for different duration categories.



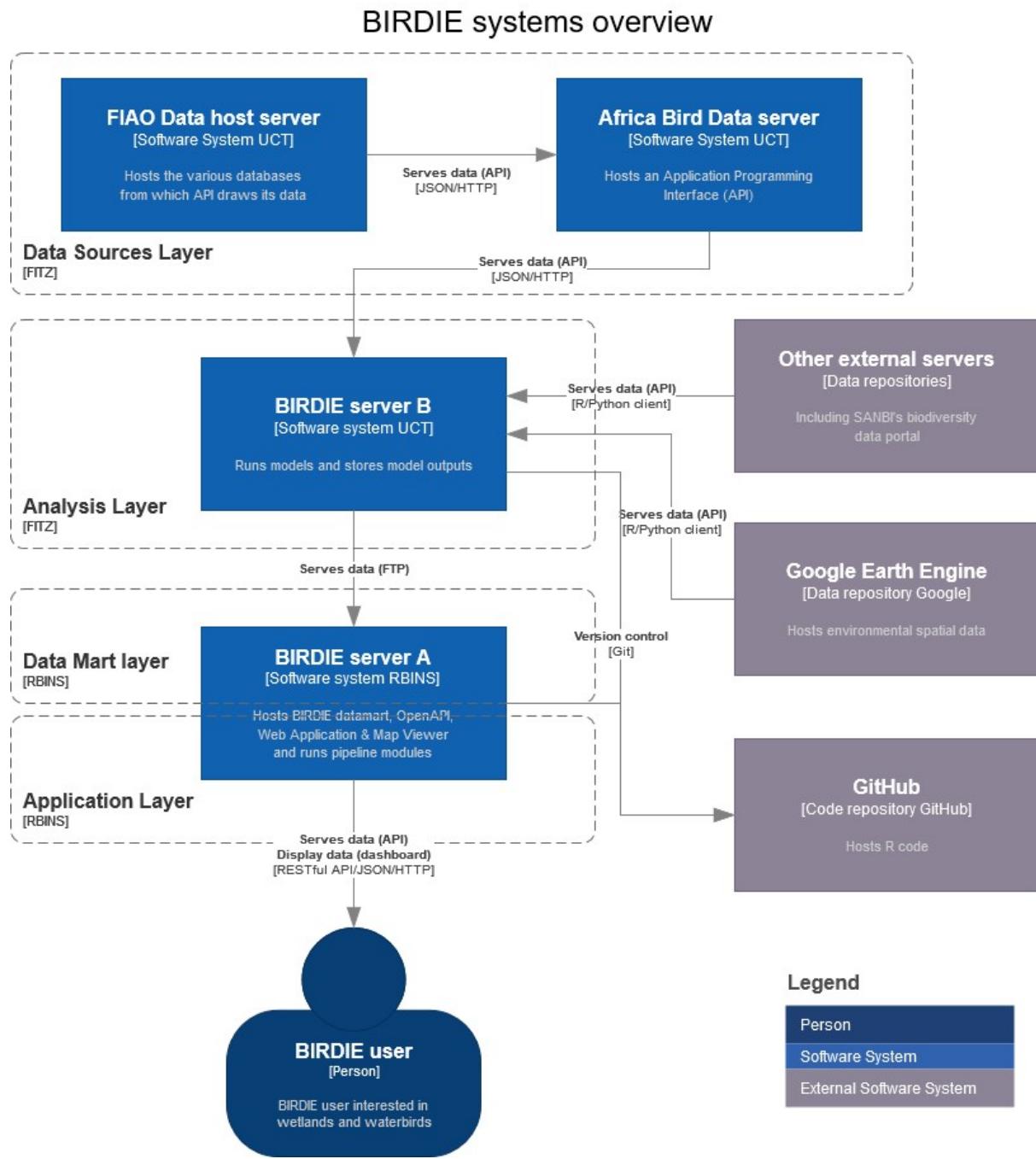
**Figure 4.** Number of SABAP2 cards recorded for the South African pentads between 2008-2021, in logarithmic scale. We can see how areas close to large cities in the Western Cape and Gauteng provinces, accumulate larger efforts. We can also appreciate sampling biased towards roads, particularly in the northwest of the country.



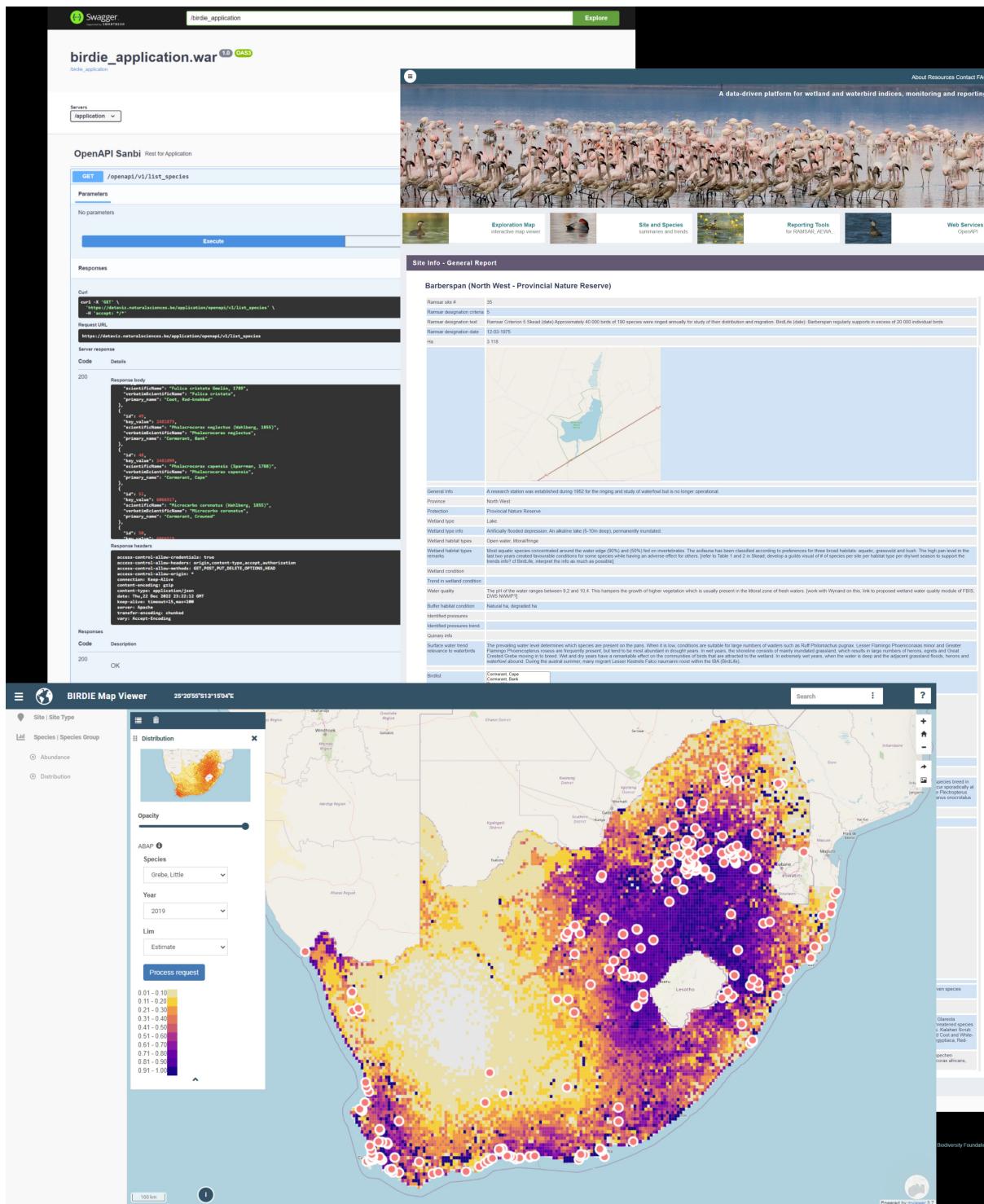
**Figure 3.** Number of SABAP2 cards recorded for the South African pentads between 2008-2021, in logarithmic scale. We can see how areas close to large cities in the Western Cape and Gauteng provinces, accumulate larger efforts. We can also appreciate sampling biased towards roads, particularly in the northwest of the country.



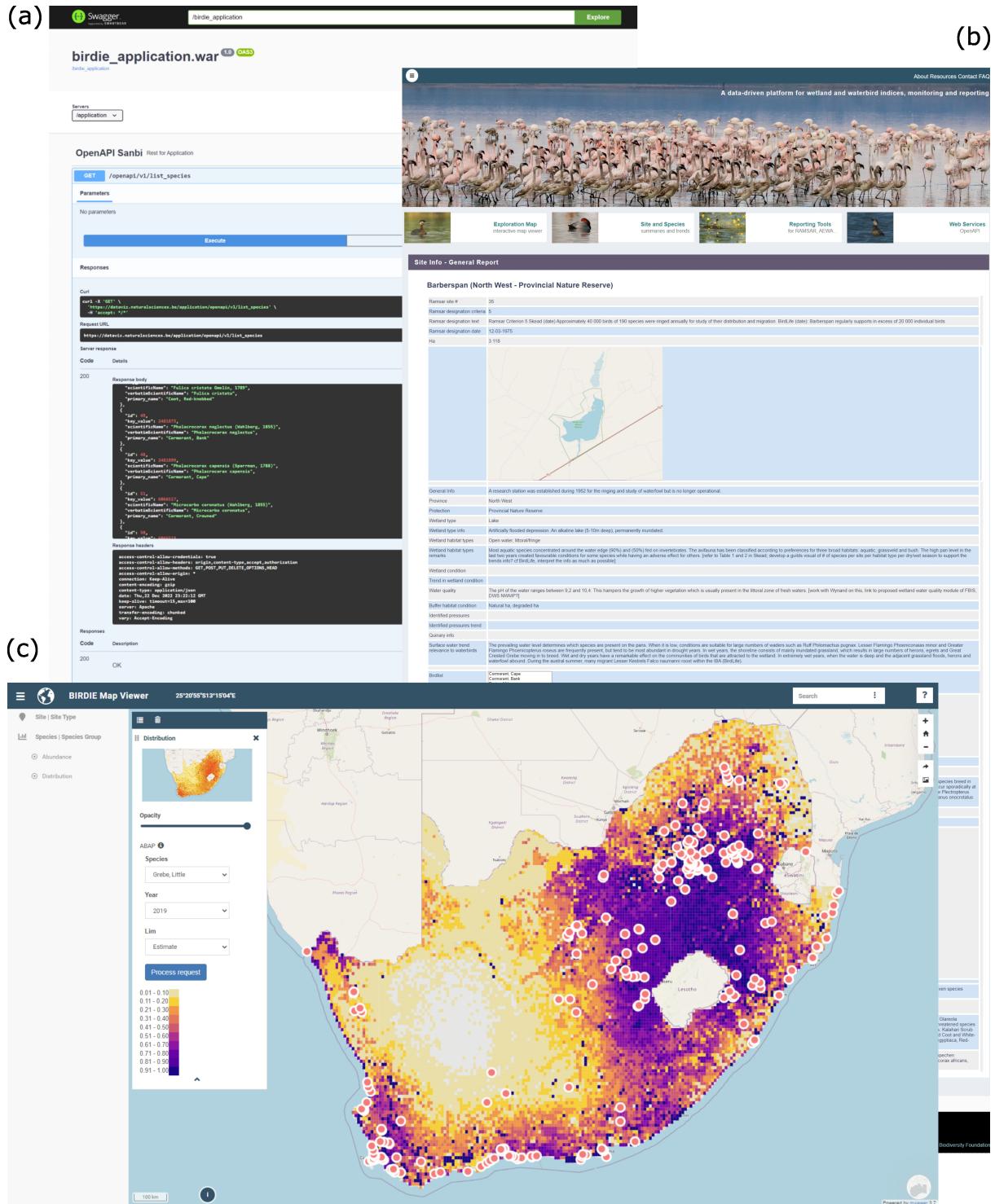
**Figure 5.** Overview of BIRDIE's server architecture. Data flows from CWAC, ABAP and other external servers into BIRDIE server B to be processed and analysed by the R modules, then these outputs move into the data mart in BIRDIE server A, which is the gateway for the dashboard and the final users.



**Figure 4.** Overview of BIRDIE's server architecture. Data flows from CWAC, ABAP and other external servers into BIRDIE server B to be processed and analysed by the R modules, then these outputs move into the data mart in BIRDIE server A, which is the gateway for the dashboard and the final users.



**Figure 6.** Basic elements of the BIRDIE web application: (a) the web services API offers a flexible framework to access the database, facilitating integration with other workflows and platforms, (b) bespoke reports for species, sites and conservation programmes and agreements such as Ramsar or AEWA, and (c) a map viewer that allows flexible exploration of the different BIRDIE indicators.



**Figure 5.** Basic elements of the BIRDIE web application: (a) the web services API offers a flexible framework to access the database, facilitating integration with other workflows and platforms, (b) bespoke reports for species, sites and conservation programmes and agreements such as Ramsar or AEWA, and (c) a map viewer that allows flexible exploration of the different BIRDIE indicators.

## TABLES

**Table 1.** Main indicators produced by the BIRDIE pipeline for waterbird species. For each indicator, we show the inputs, which can be databases (Coordinated Waterbird Counts - CWAC and the African Bird Atlas Project - ABAP), or other indicators; models used to compute the indicator (state-space model - SSM, and occupancy model - Occupancy) or whether it was computed by aggregating other lower-level indicators; the smaller spatial scale of assessment; and the smaller temporal scale of assessment. Annual changes in all of these indicators are also computed, and other indicators will be added over time as needed.

Indicator	Input	Model	Spatial scale	Temporal scale
Abundance	CWAC	SSM	CWAC site	2 seasons/year
Diversity	ABAP	Occupancy	Pentad	Annual
Extent of occurrence	Occurrence	Aggregated	National	Annual
Area of occupancy	Occurrence	Aggregated	National	Annual
Population size	Abundance	Aggregated	National	2 seasons/year
Pop. proportion on site	Abundance	Aggregated	CWAC site/national	2 seasons/year
Waterbird Conservation Value	Abundance	Aggregated	CWAC site/national	2 seasons/year
Number of sites	Abu./occur.	Aggregated	National	Annual

## TABLES

**Table 1.** Main indicators produced by the BIRDIE pipeline for waterbird species. For each indicator, we show the inputs, which can be databases (Coordinated Waterbird Counts - CWAC and the African Bird Atlas Project - ABAP), or other indicators; models used to compute the indicator (state-space model - SSM, and occupancy model - Occupancy) or whether it was computed by aggregating other lower-level indicators; the smaller spatial scale of assessment; and the smaller temporal scale of assessment. Annual changes in all of these indicators are also computed, and other indicators will be added over time as needed.

Indicator	Input	Model	Spatial scale	Temporal scale
Abundance	CWAC	SSM	CWAC site	2 seasons/year
Diversity	ABAP	Occupancy	Pentad	Annual
Extent of occurrence	Occurrence	Aggregated	National	Annual
Area of occupancy	Occurrence	Aggregated	National	Annual
Population size	Abundance	Aggregated	National	2 seasons/year
Pop. proportion on site	Abundance	Aggregated	CWAC site/national	2 seasons/year
Waterbird Conservation Value	Abundance	Aggregated	CWAC site/national	2 seasons/year
Number of sites	Abu./occur.	Aggregated	National	Annual