LOCAL ENVIRONMENT AND SAMPLING BIAS DRIVE PREVALENCE ESTIMATES IN FISH COMMUNITIES

1 | Introduction

Despite often being neglected in ecological studies (Morley, 2012; Scholz & Choudhury, 2014), parasites are key components of natural communities and ecosystem functioning (Dobson & Hudson, 1986; Frainer et al., 2018; Lafferty et al., 2008; Marcogliese, 2004; Minchella & Scott, 1991; Pascal et al., 2020; Poulin, 1999). They often involve physiological and behavioral changes of their hosts (Barber et al., 2000; Iwanowicz, 2011), are major elements of food webs (Lafferty et al., 2006, 2008; Marcogliese & Cone, 1997a) and respond to global changes (Brooks & Hoberg, 2007; Cable et al., 2017; Carlson et al., 2020; Mostowy & Engelstädter, 2010). Thus, host-parasite association dynamics are crucial to understand as parasite are ubiquitous in ecological systems (Windsor, 1998). Despite that recognition, infection patterns within a population or community can vary across taxa, time and/or space making it difficult to understand and model the drivers of infection within ecological systems (Happel, 2019; Poulin, 2006; Poulin & Dick, 2007; Thieltges & Reise, 2007; Villalba-Vasquez et al., 2018; Young & Maccoll, 2017). Consequently, our understanding of the processes explaining variation in parasite infection across spatial scales is still poor (Bolnick et al., 2020; Poulin, 1998). Studies are constraint by context-dependencies such as spatial scaling, biological scaling, and sampling design. In this context, multi-scale approach to host-parasite ecology is imperative. As humans modify ecosystems (physically and biologically) it is essential to understand the full extent of complexity of disease dynamics to improve future predictions and select appropriate scale for management and conservation strategies (Cohen et al., 2016).

The distribution of organisms in space is constrained by environmental filtering (i.e. local abiotic conditions must be ecophysiologically suitable for species to occur), species interactions and dispersal limitations (Leibold et al., 2004). Spatial scale-dependance of these processes have long been established in ecology (Rietkerk et al., 2002; Viana & Chase, 2019). In this regard, single-scale studies tend to overlook the full extent of ecological processes shaping species distribution

31   and interactions, hence drawing biased conclusions (Levin, 1992; Peterson & Parker, 1998). This

32   has also been addressed in disease ecology. For example, Cohen et al., (2016) found that the

33   relative importance of processes driving the distribution of disease emergence (host richness,

34   abiotic factors, and human population density) varied with the spatial extent of the study. Scaling-

35   level on which empirical data is scoped for estimating infection parameters can mislead

36   conclusions if cross-scale dependencies are unknown. For instance, in a study on nematodes

37   infecting alfalfa (*Medicago sativa*) at three scale levels (river basin, field and microhabitat), the

38   abundance of parasites at one scale did not explain the abundance at another scale revealing the

39   importance of cross-scale processes (e.g., dispersion by irrigation water) (Simmons et al., 2008).

40   One of the first steps to clarify these context-dependencies in infection dynamics is to address the

41   importance of spatial scaling for the estimation of infection parameters such as the prevalence of

42   infection in biotic communities. Patterns of regional-scale infection parameter can be caused by

43   different underlying local patterns of infection. For instance, a regional prevalence of 20% could be

44   caused by all sites having the same prevalence with no variance (i.e., all sites at 20%) or by the

45   average prevalence of all sites being at 20% with large inter-site variance (e.g., half at 0% and half

46   at 40%). Thus, context-dependencies might be the reason why infection patterns are still

47   misunderstood across scales.

48

49   Scaling is especially relevant in parasite ecology given the hierarchical organization of both

50   the parasite and host component. As an example, the simplest clustering organization of parasites

51   is called an infrapopulation (a parasite population within a host individual) but parasite species can

52   also co-exist with other parasite species within an individual host (infracommunity) or within a host

53   population (component community) (see Bush et al., 1997 for more extended terminology).

54   Considering the biological organizational complexity of host-parasite associations, there are still

55   many biological scale combinations (consisting of a parasite and a host element) that need to be

56   investigated. For instance, component communities of metazoan parasites in fish host-populations

57   have been the focus of many studies in the past decades regarding spatiotemporal patterns (e.g.

58   Carney & Dick, 2000; González & Poulin, 2005; Pérez-del-Olmo et al., 2009; Villalba-Vasquez et al.,

59   2018). For exemple, Villalba-Vasquez et al. (2018) looked at infracommunities and component

60   communities of parasites in Panama spadefish (*Parapsettus panamensis*) collected over 6 years

61   from 4 locations and found similar patterns of low species richness, low diversity and dominance

62   from a single parasite species across locations, but high variations in species composition and

63   community structure between years and locations. Conversely, species-level infection patterns in

64   host-communities (i.e., one parasite species infecting many host species in a community) remains

65   largely unexplored across spatial scales (but see Bielby et al., 2021; Moore et al., 2014; Olsson-

66   Pons et al., 2015). In an ecological perspective, studying infections at the host-community level is

67   coherent because hosts can co-exist within the same habitat exposing them to similar encounter

68   chances with infectious stages of parasites. Besides, a study on macaques found that omitting

69   individuals of a social network can change the strength of the relationship between parasites

70   intensity and social network centrality (Xu et al., 2022), suggesting that the interactions between

71   hosts of a population or community are a structuring factor of parasite infection.

72

73   Despite recent efforts, it remains unclear whether infection patterns result from

74   stochasticity (i.e., infection patterns do not differ from randomness) or determinist filtering (i.e.,

75   infection patterns are explained by a set of known drivers and mechanisms) (e.g., Carney & Dick,

76   2000; González & Poulin, 2005; Kennedy, 2009; Poulin & Valtonen, 2002). Studies have shown that

77   both local biotic (e.g., parasite-parasite interactions, host behavior, feeding preference) and abiotic

78   (e.g., habitat structure, water chemistry, pollution) components can be correlated to variation in

79   infection metrics at various spatial scales (Altman & Byers, 2014; Falke & Preston, 2021; Lagrue et

80   al., 2011; Lagrue & Poulin, 2015; Poulin & Morand, 1999; Thieltges et al., 2008). Individual host

81   characteristics such as sex (Zuk & McKean, 1996), age/size (Marcogliese et al., 2001; Poulin, 2000)

82   and genetics (Williams-Blangero et al., 2012) drive host susceptibility leading to sources of variation

83   in infection parameters of host-populations. For example, Poulin (1996) found that infection

84   prevalence of trematodes was higher in male individuals across bird and mammal hosts. Infection

85   prevalence can thus be sex-biased within host-populations (Kowalski et al., 2015). On a larger scale,

86   host community properties such as species richness and host population density can create a

87   "dilution effect" by reducing a parasite's encounter rates with hosts (Ahn & Goater, 2021; Buck &

88   Lutterschmidt, 2017; Civitello et al., 2015; Dargent et al., 2013; Lagrue & Poulin, 2015). For

89   instance, alternative host species can act as a decoy deceiving infection on a focal host species

90   (Ahn & Goater, 2021). Local habitat characteristics have also been correlated to variation in

91   parasitological parameters in host communities. In American eels (*Anguilla rostrata*), parasite

92   diversity decreases when water pH is below 5.4, with some parasite families such as digenean

93   trematodes being absent below pH 4.7 (Marcogliese & Cone, 1997b). Similarly, lentic water bodies

94   with dense vegetation display higher prevalence and abundance of *Posthodiplostomum cuticola*

95   parasites in host fish than rivers and reservoirs with steep banks (Ondrackova et al., 2004),

96   suggesting that local habitat and morphometry of the water body are drivers of encounter filtering

97   (sptatiotemporal window that allows encounter between the parasite and its host). Furthermore,

98   spatial features of an ecosystem (e.g., watershed, connectivity) act as large scale filtering by

99   limiting dispersal of both parasites and hosts (Bolnick et al., 2020) thus creating infection

100   patchiness at the landscape level. Happel (2019) provided evidence for consistent geographical

101   pattern in the prevalence of trematodes between watersheds. Bolnick et al. (2020) also found that

102   increased waterways distance between waterbodies lead to more different parasite community's

103   composition in threespine stickleback (*Gasterosteus aculeatus*). Waterbodies such as lakes are

104   therefore good natural study systems to investigate scaling effect as they are delimited entities,

105   connected or not in space through streams. Investigating a wide range of predictors has the

106   potential to enlighten processes driving infection patchiness (i.e., infection hotspots) within a

107   system that appears homogenous (Bolnick et al., 2020). Consequently, a complete ecological

108   perspective on host-parasite systems must incorporate abiotic, biotic and spatial ecosystem

109   elements to highlight mechanisms shaping patterns of infection as some filters are scale-

110   dependent (Bolnick et al., 2020; Cohen et al., 2016).

111

112   Relevance of sampling design has been long established for empirical studies on wild

113   populations. Researchers working on wildlife must deal with a tradeoff between minimizing stress

114   of focal species and impact of natural habitats while generating accurate estimates of populations

115   and communities when designing their sampling protocols. Precision of population parameters

116   (e.g., abundance, recruitment, age classes, sex-ratio) can also be influenced by the sampling

117   method and/or the sampling effort (i.e., the number of sampling). For example, Khaemba et al.,

118   (2001) simulated three sampling designs based on empirical data and showed that the intensity of

119   sampling (here the closeness of transects) improved precision of the estimated population in all

120   three designs. Sampling methods can introduce bias into estimates based on animal personality

121   and behavior (Biro & Dingemanse, 2009). Animal personality is often described on a shy-bold

122   continuum that influences risk-driven decision (Wilson et al., 1993), on what relies on a lot of

123   detection methods used in the field. For example, a study on the black rat (*Ruttus rattus*) showed

124   that sampling devices that required a higher level of intimacy detected fewer individuals, as the

125   perceived risk gets higher the closer they got to the traps thus creating a "personality filtering"

126   favoring sampling of bold individuals (Johnstone et al., 2021). Trappability in aquatic habitats follow

127   the same pattern. In Biro & Post, (2008), bold rainbow trout were caught three times as often as

128   shy individuals in gillnets. Thus, estimations based on empirical surveys are context-dependent on

129   how the methodology used is personality-biased. Although it has not thoroughly been investigated

130   yet, there is no reason indicating different outcomes for estimates of infection parameters in the

131   wild. A study on the chronic wasting disease in deer species showed that the survey method used

132   to estimate the disease prevalence can lead to overestimation or underestimation of the true

133   infection parameter (Nusser et al., 2008). Wilson et al. (1993) also observed different infection

134   levels in Pumpkinseed sunfish (*Lepomis gibbosus*) depending on the fishing method. Poulin (2013)

135   found a weak correlation between the aggregation level of helminth parasites in fish hosts and the

136   number of individual-host sampled (sample size). Consequently, sampling design elements must

137   be considered when studying scale-dependencies of infection parameters.

138

139   Here, we explore the effect of spatial scale on infection prevalence estimates across

140   different sampling methods using data on the black spot disease in littoral fish communities across

141   15 lakes varying in morphometric attributes and local biotic and abiotic conditions. We analyze

142   infection prevalence data at the host-community level (all fish hosts) for scaling-dependencies

143   through three spatial scale levels (landscape-, lake- and site-scale) in order to investigate (i) the

144   effect of increasing sampling effort on landscape prevalence estimation (random resampling

145   accumulation curves) through different sampling methods (ii) the distribution of observed lake

146   prevalence estimates across the landscape and (iii) the importance of ecological predictors on

147   infection prevalence on a fine-scale sampling. For landscape-scale analysis, we generated random

148   prevalence accumulation curves by resampling across all lakes. This approach allows us to compare

149   regional-scale prevalence estimates among different sampling methods along a gradient of

150   increasing sampling effort (or area sampled) in order to evaluate the number of samples needed

151   to correctly estimate the landscape prevalence (Figure 4.A). For lake-scale analysis, we used a

152   frequency distribution of observed infection prevalence in lakes to better understand the infection

153   dynamics across the landscape (Hess et al., 2002; McGeoch & Gaston, 2002)(Figure 4.B). The final

154   step was to test a suit of deterministic environmental predictors underlying different scale-

155   dependent processes to understand the observed site-scale infection prevalence. The shapes and

156   strength of the relationships between infection parameters and drivers are a key element to better

157   understand the mechanisms behind infection patterns (Poulin et al., 2011).
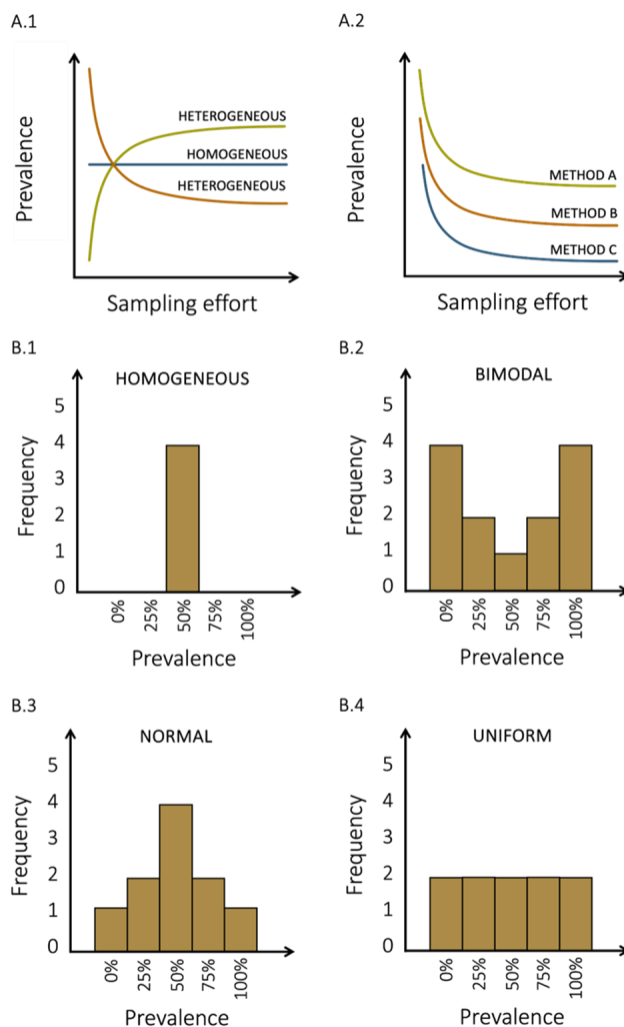
158

159

160

Figure 4. Case scenarios of sampling effort effect on landscape prevalence estimates and frequency distributions of lake prevalence estimates. A.1) Prevalence could be homogeneous (blue line) or heterogeneous (green and brown lines) across the landscape. In heterogeneous scenarios, prevalence is either overestimated (brown line) or under underestimated (green line) before stabilizing around the accurate prevalence. A.2) Sampling methods could lead to different prevalence estimates across the landscape. Height difference between lines would represent method biases. B.1) All lakes could be infected at the same prevalence level. B.2) Lake's frequency distribution could follow a bimodal pattern where low and high prevalence are more common across the landscape. B.3) Lake's frequency distribution could follow a normal pattern where mean landscape prevalence is more common. B.4) Lake's frequency distribution could follow a uniform pattern where all infection levels are equally represented.

2 | Methods

2.1. Host-parasite system

Black spot disease is a common fish infection in marine and freshwater ecosystems caused by Digenean trematodes (flukes) (Kurochkin & Biserova, 1996). While some aspects of the ecology of some of these parasite species (e.g., *Posthodiplostomum cuticola, Uvulifer ambloplitis, Crassiphiala bulboglossa* and *Apophallus brevis*) are relatively well studied, spatiotemporal infection patterns and their relation to environmental conditions remain poorly understood. Black spot-causing trematodes all have a similar complex life cycle requiring snails, fish and piscivorous birds as hosts (Figure 4). The parasite encysts under the skin, in the fins or the muscles of the fish host (Hoffman, 1956; Krull, 1932, 1934) where it can survive in a dormant form for a few years (Hoffman & Putz, 1965) until the infected fish gets eaten by a bird. The distinctive black cyst symptom results from melanin pigment stimulation in the fish tissues induced by trematode cercaria penetration (Davis, 1967; Lemly & Esch, 1984a). In the context of this study, we will refer to the black spot-causing trematodes as a guild because they have similar ecological requirements and cause the same symptoms in fishes.

Akin to other parasitic diseases, the infection parameters of the black spot disease can vary across time, space, and species. For example, the infection prevalence in bluegill sunfish (*Lepomis macrochirus*) inhabiting ponds is typically at its lowest around April/May and its maximum around September/October because cercariae are released through the summer and heavily infected fishes might be killed overwinter more often than less infected conspecifics (Lemly & Esch, 1984a, 1984b). Habitat characteristics that favor snail populations such as shallow waters, low water flow and abundant macrophytes have also been associated with higher abundance dans prevalence infection in fish-hosts (Marcogliese et al., 2001; Ondrackova et al., 2004). Hence, infection parameters of black spot disease in fish-hosts are usually higher in lentic systems and in the littoral zone rather than the pelagic zone (Ondrackova et al., 2004) because of a higher encounter probability. Since the black spot disease is caused by a parasite guild, a large range of fish species

190   can be infected. However, some species tend to be more heavily infected than others. For instance,

191   Lemly & Esch (1984) found that bluegill sunfish were more heavily infected than largemouth bass

192   (*Micropterus salmoides*).

193

194   *2.2. Study area*

195

196   We sampled 15 lakes from six watersheds around Saint-Hippolyte, QC, Canada (45.92°N,

197   73.97°W) (Figure 5). The lakes Corriveau, Croche, Cromwell and Triton are in the protected area of

198   the biology station of the Université de Montréal while the other lakes are surrounded by private

199   properties. Considering their geographical proximity, we assumed that all the lakes are exposed to

200   the same climatic conditions. The lakes were selected nonrandomly according to their accessibility

201   and the availability of morphometric data (see Table S1 for the morphometric characteristics).

202

203   *2.3. Data acquisition*

204   Infection prevalence was estimated as the number of individuals infected with black spot

205   disease divided by the total abundance of individuals of a given fish community. Three sampling

206   methods (minnow traps, seine nets and transects) were used to assess prevalence infection and

207   maximize the chances of catching different fish species. The field work took place from mid-June

208   to end of August 2022 and was restricted to days without precipitation to limit sampling bias due

209   to meteorological effects. Animal handling was approved by the Université de Montréal's animal

210   care committee (protocol number 22-025) and the scientific fishing permit was granted by the

211   Ministère de l'Environnement, de la Lutte contre les changements climatiques, de la Faune et des

212   Parcs (MELCCFP) of Québec (2022-05-16-1971-15-S-P).

213   Sampling effort within lakes was determined according to the lake surface area, except for

214   minnow traps (see Table S2) as setting these traps requires less time and manipulation than the

215   other methods. Seine sampling varied between four and eight efforts while the number of transect

216   sites varied between two and six per lake. 15 minnow traps were set in each lake to maximize the

217   number of captures.

218   The minnow trap and seine net samplings were both conducted on the same day for each

219   lake. The seine samplings occurred during daytime in habitats lacking large obstacles like trunks

220   and rocks in which the net can get stuck. Two sizes of minnow traps were set at approximately

221   equal distance along the shore at 4PM and pulled four hours later, to target species that are most

222   active at dusk. All traps sat on the bottom substrate, but there were no other habitat restrictions

223   for the minnow trap placement. Half of the traps (3 large and 5 small traps) were baited with bread

224   to sample various feeding preferences and behavioral habits. The fishing gear dimensions are

225   available in Table S3. All the fishing gear was pressure washed and sun dried between lakes

226   following MELCCFP recommendations to prevent exotic species contamination.

227   All captured fishes were counted, identified to the species level, and measured (estimation

228   of the total length to the nearest centimeter) directly after capture and subsequently released.

229   Northern redbelly dace (*Chrosomus eos*) and finescale dace (*Chrosomus neogaeus*) individuals

230   were identified as *Chrosomus* spp. as the two species hybridize in these lakes and cannot be

231   distinguished based on morphology in the field (Leung et al., 2017). The presence of black cysts

232   was assessed by examining the left side of the fish (De Bonville, in prep.). Juveniles and adults were

233   included in this study as both life stages are vulnerable to black spot disease.

234   Fifty-meter snorkeling transects along the shore were conducted to assess black spot

235   infection prevalence in the fish communities. Due to time constraints, lakes Beaver, Montaubois,

236   Tracy and St-Onge were excluded from the transect sampling because of low fish abundance in the

237   fishing samples or poor water transparency. We selected sites that were approximately between

238   0.5 and 3.0 m deep, not fully covered by macrophytes, not obstructed by docks and preferably

239   with some vegetation, rock, or trunk refuges for the fishes. The sites were selected and flagged

240   every ten meters at the end-May 2022. The transects were performed simultaneously by two

241   observers each monitoring one-meter field of view in front of them and moving forward at a pace

242   of three minutes per ten meters. The fishes coming in the field of view from behind the observer

243   were not counted. All the fishes were identified to the lowest taxonomic level possible and visible

244   black spot infection was noted. All cyprinids were grouped into the same category since they move

245   fast and are more difficult to identify at the species level in these circumstances.

246   Following the prevalence transect assessment, we completed the habitat description. The
247   percentage of four substrate categories (silt, sand, rock, and boulder), the macrophyte coverage,
248   the number of trunks (or large branches), and the mean depth was estimated for each 10-meter
249   transect section. The mean estimate of each category was then calculated for all variables at the
250   transect-scale. Temperature, dissolved oxygen, conductivity and pH were measured at mid-depth
251   (around 0.3 to 2 meters) at the beginning of every transect with a YSI ProDSS Multi-Parameter
252   Water Quality Meter. 1 L of unfiltered water was taken at mid-depth in previously acid-washed
253   HDPE bottle for carbon and nutrient content analyses. The samples were placed in a dark cooler
254   and transported back to the field station laboratory where the 1 L samples were separated in
255   previously acid-washed 40 mL vials to measure total organic carbon (TOC), and 500 mL HDPE bottle
256   for total nitrogen (TN) and total phosphorus (TP). TOC samples were placed in a 4°C refrigerator
257   and analyzed within the week while TN and TP samples were kept in a -20°C freezer until later
258   processing. TOC sample contents were measured on a Siervers M5310 C Laboratory Total Organic
259   Carbon Analyzer. TN and TP samples were oxidized with persulfate and autoclaved the day before
260   analysis following EPA methods 353.2 and 365.3 respectively (US EPA, 1978, 1993). TN
261   concentrations were measured on a Lachat QuikChem 8500 analyzer and TP concentrations were
262   analyzed on a Asoria-Pacific Astoria2.
263
264   *2.4. Statistical analyses*
265
266   We employed the following procedures to describe the black spot prevalence pattern across
267   multiple scales (landscape, lake and site). All data manipulations and analyses were conducted
268   using R (version 4.2.3). Lake Tracy was excluded from the lake-scale analysis because of low fish
269   abundance data obtained with the fishing methods (N = 1 individual sampled).
270
271   For the landscape-scale analysis, we looked at the effect of the sampling design (here
272   sampling method and sampling effort) on the estimation of landscape prevalence. We used a
273   resampling approach on the data from the different sampling methods (minnow trap, seine net,
274   transect, and all methods combined) to investigate how the infection prevalence estimate changes
275   with an increasing sampling effort gradient. The sampling effort prevalence was calculated as the

276   number of infected individuals divided by the total number of individuals regardless of the lake

277   sampled. For each sampling method, we randomly selected N number of samples from our data to

278   estimate the landscape (weighted mean) prevalence. The operation was repeated 999 times for a

279   sampling gradient from 1 to 35 samples (the smallest landscape-scale sampling effort being 39 for

280   the transect method, we did not want to sample the entire data set). A local regression was then

281   fitted to the landscape estimates generated by our simulation of each sampling effort value (N) for

282   visualization of the tendency through the sampling gradient. We estimated the final landscape

283   prevalence and the approximate sampling effort needed to reach a steady prevalence value for

284   each method.

285

286   For the lake-scale analysis, we examined the frequency distribution of the lakes' prevalence

287   according to the sampling method. The mean infection prevalence was previously calculated for

288   each lake, and visualization was made with histograms set at six bins to avoid gaps. The distributions

289   were then compared with the map of the study area for spatial patterns.

290

291   For the site-scale analysis, we modeled the relationships between environmental drivers

292   and the community infection prevalence on a transect level. We used the prevalence data from the

293   transect method because it is the only method that allows an accurate association of infection

294   prevalence with all environmental variables (including those measured at the site-scale), thus can

295   best reveal the key drivers of infection prevalence. Prior to analysis, we explored the data for

296   outliers in both the response and explanatory variables, for collinearity between explanatory

297   variables, and for non-linearity in the relationships between the response and the explanatory

298   variables following recommendations by Zuur et al. (2009). Our data showed some non-linear

299   patterns. Consequently, we used generalized additive models in a mixed-model framework

300   (GAMMs) fitted using marginal likelihood (ML) estimation to account for spatial clustering of sites

301   from the same lakes. We used a quasi-binomial family distribution because our response variable

302   (infection prevalence) is a proportion constrained between 0 and 1 and, to account for

303   overdispersion in the data. The optimal amount of smoothing (λ) was internally determined by the

304   modeling function from the *mgcv* package (Wood, 2017). Since we had a small sample size for the

305 transect method (N = 39), we modeled one environmental variable at the time with a cubic spline
306 and a random effect on the lake variable to save degrees of freedom. The deviance explained (D^2)
307 was used as a measure of the model fit. The model validations were conducted with *gratia* package
308 (Simpson, 2023). The visualization of the partial effects was made with *gratia* (Simpson, 2023) and
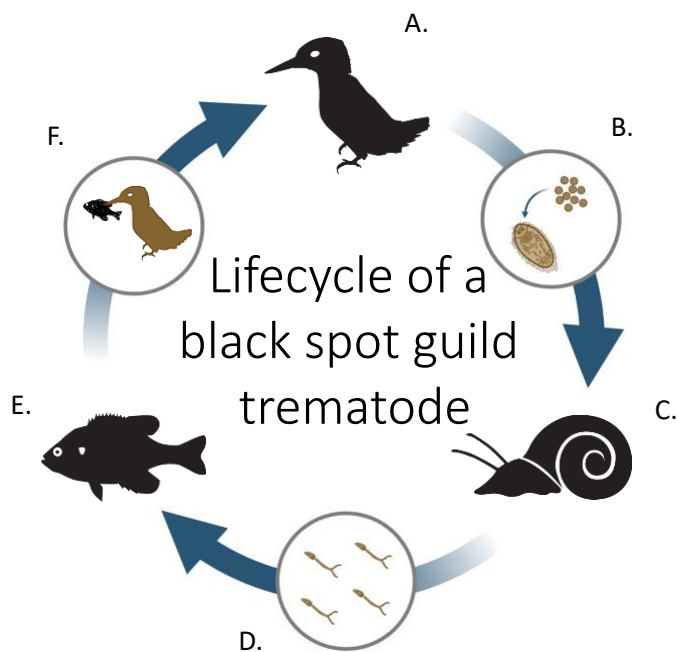309 *ggplot*2 (Wickham, 2016) packages.
310

Figure 5. The typical lifecycle of a parasite species causing the black spot disease in freshwater fishes. **(A)** The definitive host is a piscivorous bird, usually a kingfisher or a great blue heron. The metacercariae mature to its adult form and reproduce in the bird digestive system. **(B)** The eggs are released in the water through the bird feces and grow into miracidia (first larval form). **(C)** The miracidia infect the first intermediate host, a freshwater snail of *Helisoma* or *Planorbella* genus. The larvae develop into free-living stage cercariae (second larval form) and leave the snail. **(D)** The cercariae swim until they encounter a potential fish host. **(E)** The second intermediate host is a fish. The cercariae penetrate under the skin or the fins and transform in the metacercaria form. **(F)** The fish is eaten by a definitive host.

3 | Results

Our main goal here was not to investigate how prevalence of black spot disease in fish communities is perceived across three scale levels (landscape, lake, site). First, we built landscape-level infection prevalence accumulation curves, then we compared the frequency distribution of the lake's prevalence for each sampling method, and lastly, we identified predictors of site-scale prevalence patterns.

We recorded a total of 11 297 individual fishes belonging to 15 species for this study (Table S4). The minnow traps caught 1906 individuals from 10 species while seine nets caught 2427 individuals from 14 species (Table S5 and S6). 6964 individuals belonging to five taxonomic groups (four species and one family) were observed in the snorkeling transects (Table S7). The mean length of all fish captured through fishing methods was 5.59 ± 2.96 cm (N = 4333). The overall mean length of each lake, species and, species within each lake are presented in Table S8, S9 and S10 respectively.

All fish species sampled, except for *Ameiurus nebulosus* and *Esox masquinongy* are susceptible to black spot disease (see Table S11). However, our survey only found evidence of black spot infection in pumpkinseed sunfish (*Lepomis gibbosus*), rock bass (*Ambloplites rupestris*), yellow perch (*Perca flavescens*), smallmouth bass (*Micropterus dolomieu*), creek chub (*Semotilus atromaculatus*) and fathead minnow (*Pimephales promelas*). Details on infection prevalence of fish communities are presented for every sampling method in Table S12 (landscape-scale), Table S13 (lake-scale) and Table S14 (site-scale).

3.1. Landscape-scale

We used a resampling approach to compare how different sampling methods change the mean prevalence estimate through an increasing sampling effort across the landscape (Figure 7).
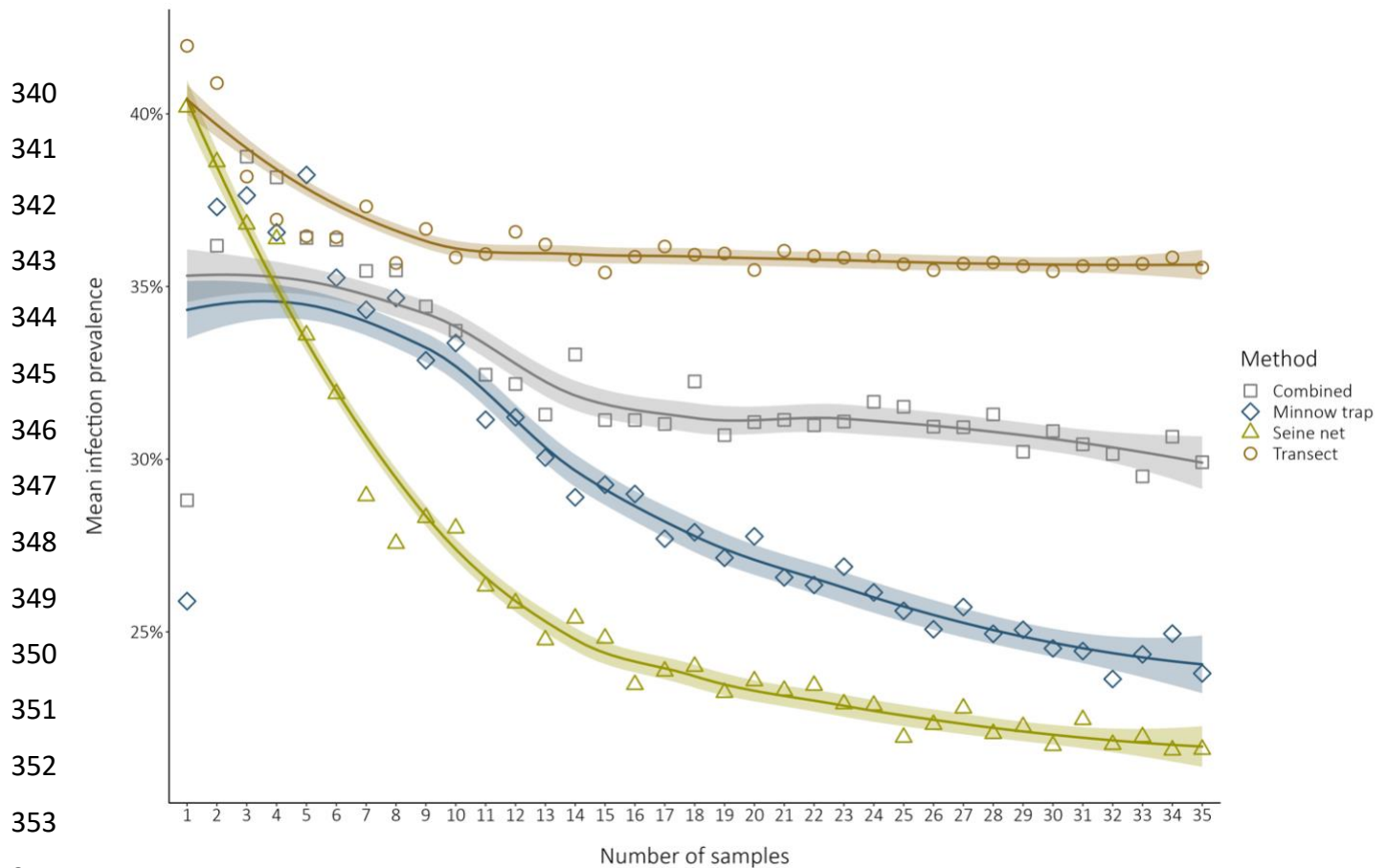
Figure 7. Method comparison of landscape prevalence estimates through an increasing random sampling effort. The data points indicate the mean prevalence for a given number of samples (N) randomly selected. The shadings indicate the 95% confidence interval on the loess smooth. The pool data used for the resampling with the minnow traps, the seine nets, the transects and all methods combined respectively contained 225, 75, 39 and 339 samples.

Each method curves overestimated the mean prevalence at low N values (number of samples) suggesting that infection prevalence is heterogeneously distributed in our landscape with presence of infection clusters or hotspots. The estimate of landscape prevalence differed between the sampling methods, varying between 21% and 36%. After 35 samples, the transect method generated the highest mean prevalence (36 %) followed by the methods combination (31 %), the minnow trap method (24 %) and the seine method (21 %) (see Table S15). However, the minnow trap curve did not stabilize after the 35 random samples suggesting that the prevalence value obtained is higher than the actual estimate (landscape infection prevalence measured by the minnow trap is 19%, see Table S12). However, relatively few random samples are necessary to estimate a landscape prevalence for the transect method (approximately 10 samples) while an

369 accurate prevalence estimate only occurs after 30 samples for the seine method. Even if the
370 method combination curve (in gray) stabilizes around 20 samples, it nevertheless displays some
371 variation around the curve because of the variability among the different methods used.

372

373 *3.2. Lake-scale*

374

375 We compared the frequency distribution of the lake's prevalence for each sampling method
376 to see how the landscape prevalence is distributed among lakes and in their geographical context
377 to identify spatial patterns (Figure 6).

378

379 Following the landscape results, the lake fish community prevalence estimates are not
380 constant across the landscape. Both frequency distributions and prevalence maps support a
381 difference between methods in terms of sampling fish communities and/or behavior. The lake
382 prevalence frequency distributions for the combined methods and for the transect method show
383 a bimodal distribution, with the landscape composed of many low-prevalence and high-prevalence
384 lakes. However, neither method yielded prevalence estimates over 80%. The distribution patterns
385 for the two fishing methods (seine net and minnow trap) are less clear. These methods show a
386 right-skew distribution representing more low-prevalence lakes. Accordingly, seine and minnow
387 traps captured very similar prevalence estimates at the landscape-scale (20.4% and 19.2%
388 respectively). Heavily infected and less infected lakes do not appear to be clustered in space at the
389 regional scale (see maps in Figure 6). Moreover, close and connected lakes do not appear to follow
390 a spatial infection gradient, suggesting that geographic attributes (e.g., position in the landscape,
391 connectivity, distance to nearest lake, belonging watershed) are not important drivers of the local
392 infection prevalence. Nevertheless, lake-scale geographic patterns were not statistically tested.
393 Both frequency distributions and prevalence maps support a difference between methods in terms
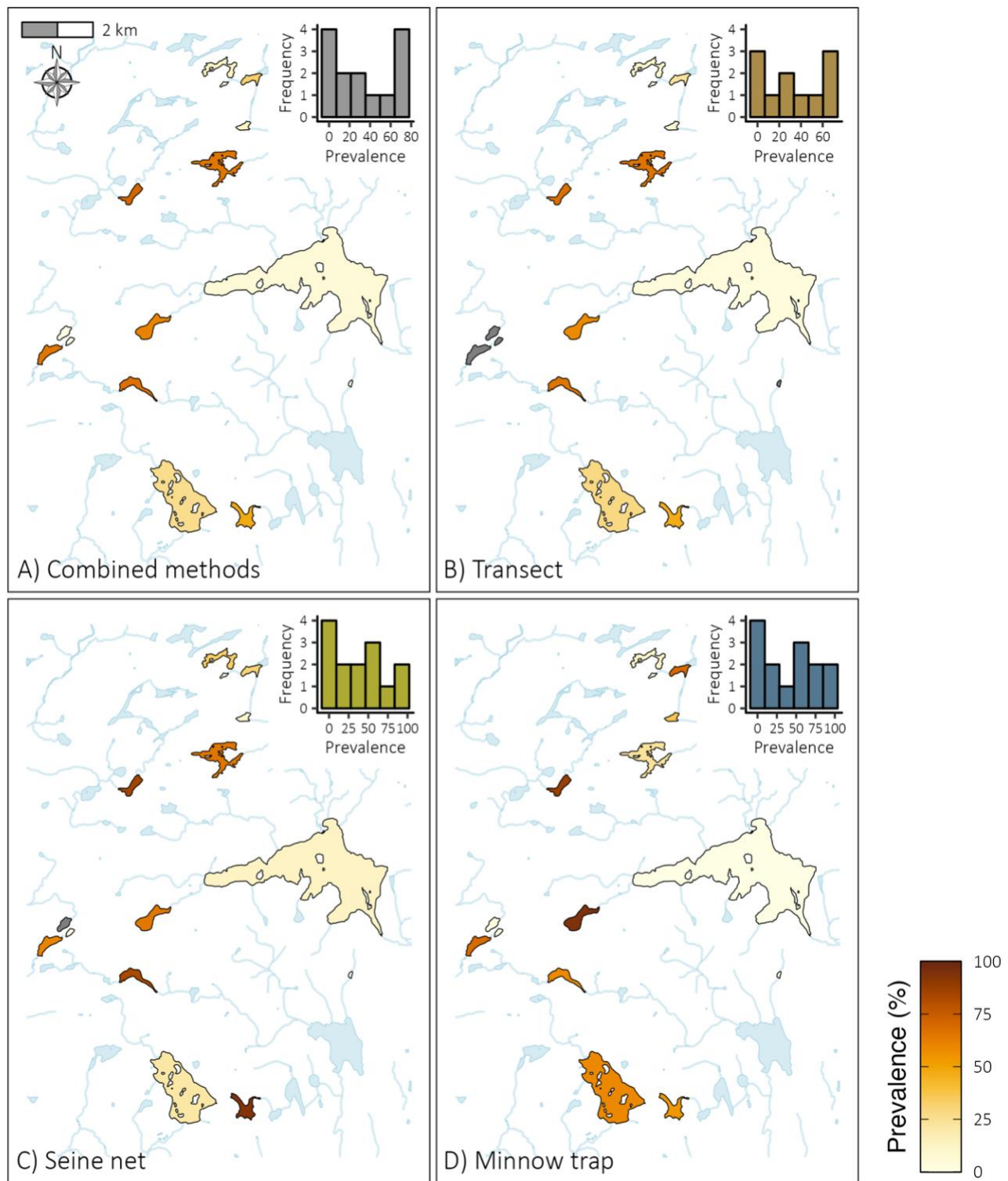394 of sampling fish communities and/or behavior.

395

Figure 6. Method comparison of the lakes' local community prevalence estimates across the landscape. Frequency distributions are given in the top right corner of each map. Lakes in grey were not sampled except for lake Tracy that was excluded because of insufficient data (one fish sampled through all methods). **(A)** All the sampling methods combined (weighted mean regional prevalence is 29.55 %). **(B)** The transect method (weighted mean regional prevalence is 35.55 %). **(C)** The seine nets (weighted mean regional prevalence is 20.44 %). **(D)** The minnow traps (weighted mean regional prevalence is 19.20 %).

3.2. Site-scale

The relationships between the potential predictors and the site-scale prevalence (transect-level prevalence) were assessed with generalized additive mixed effects models (see Table S16 for models' estimates). Environmental characteristics of transect sites used as model predictors are presented in Table S17.
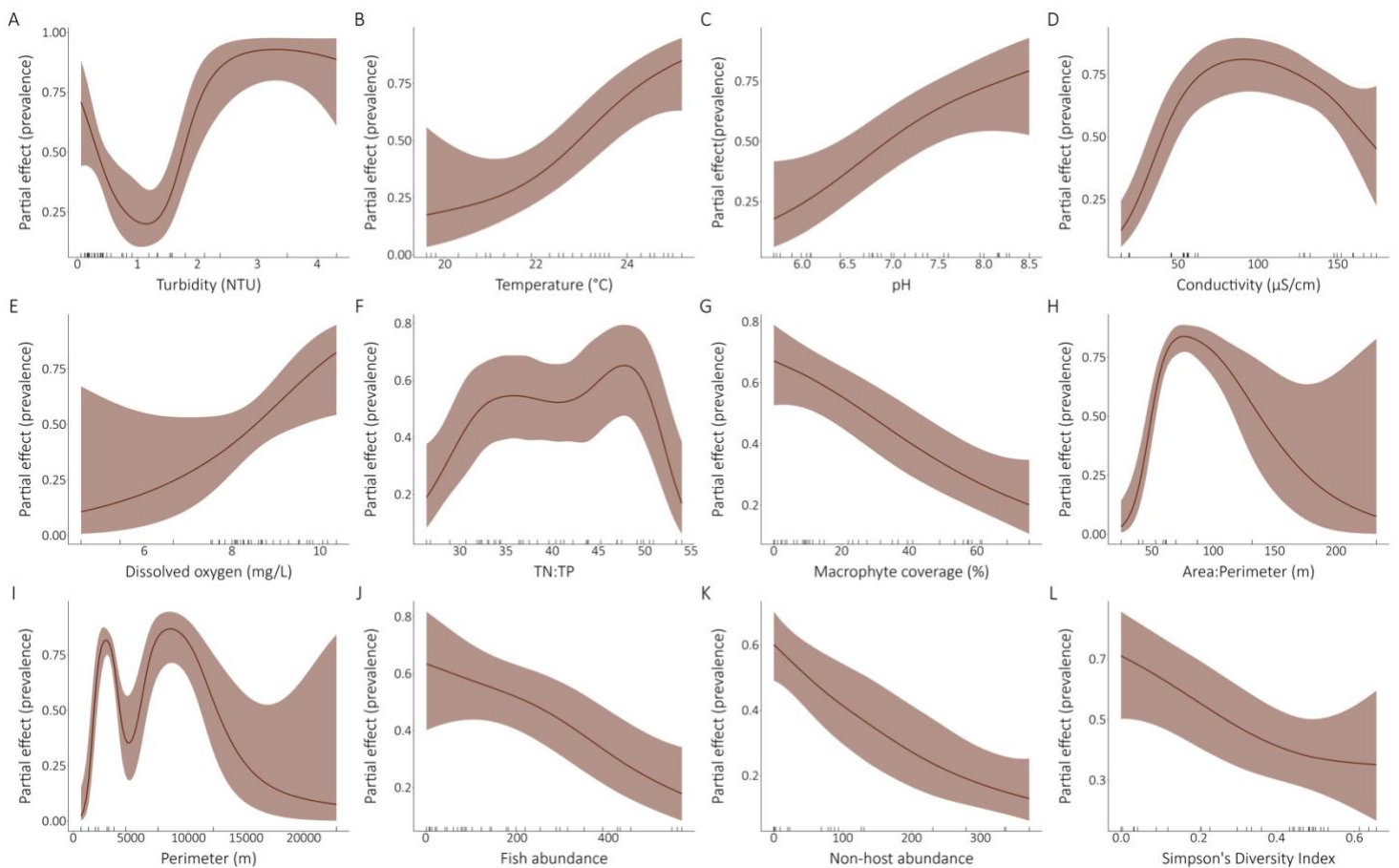

Figure 8. Relationships between the site-scale community infection prevalence and selected environmental drivers. All the models presented are univariate binomial generalized additive mixed models with the lake as a random effect on the intercept. The partial effects of the environmental variables on the prevalence are presented for the significant models. The ticks on the x-axes indicate a data point. **(A)** Turbidity. **(B)** Temperature. **(C)** pH. **(D)** Conductivity. **(E)** Dissolved oxygen. **(F)** TN:TP ratio. **(G)** Macrophyte coverage. **(H)** Area:Perimeter ratio. **(I)** Perimeter. **(J)** Fish abundance. **(K)** Non-host abundance. **(L)** Simpson's Diversity Index. Area:Perimeter ration and Perimeter are lake attributes while other parameter are site-scale measurements.

The partial effects of environmental variables on prevalence are shown for all significant models in . The turbidity model has the best fit (D^2 = 88,71%) and is mostly non-linear. The relation evidence a plateau for high-prevalence values, indicating that a prevalence saturation is reached for turbidity values above 2 NTU. That said, this model must be carefully interpreted as we sampled only a few high-turbidity sites. Water temperature has an increasingly proportional relationship with infection prevalence. The prevalence estimates also increased proportionally with pH, so that prevalence increases in more alkaline lakes. The conductivity and prevalence have a non-linear, unimodal relationship peaking around 80 (µS/cm). The relationship must also be carefully interpreted because of some gaps in the conductivity values (between 61.3 and 129.2 µS/cm). Prevalence values increases almost proportionally with dissolved oxygen concentration in the water. However, we must take into consideration that the variation interval is very large for low-concentration oxygen values because of only three values below 7 mg/L were recorded. The relationship between TN:TP ratio and prevalence is unimodal but highly non-linear. For the macrophyte coverage, we found a decreasing relationship between the amount of macrophyte cover and the prevalence of infection, meaning that low macrophyte cover correlates with high prevalence of site-scale communities. A parabolic curve is observed in the Area:Perimeter model although, high ratio values are more uncommon increasing the variation interval. The perimeter model shows a highly non-linear presenting two distinctive peaks and a large confidence interval at higher perimeter values. Total fish abundance and non-host abundance (here abundance of cyprinids) both showed a negative relationship with prevalence suggesting that sites with more fishes and/or non-host individuals have lower prevalence of infection. The relationship between the prevalence and the species diversity index shows a decreasing trend. At the site-scale, fish communities are slightly to moderately diverse (Simpson's diversity index between 0 and 0.64) indicating the dominance of some species. Since cyprinids were not differentiated, it is hard to unravel who is the most dominant species based on the transect method data.

The models for TN, TP, TOC, type de substrate (silt, sand, rock and boulder), mean transect depth, number of trunks, lake surface area, lake maximum depth, lake mean depth, distance to nearest lake, water residence time, drainage area, lake elevation, distance to the nearest lake and

species richness were not significant (see Table S18). We did not find any relationships between black spot prevalence and any spatial attributes. Water physico-chemistry and community metrics were good predictors of infection prevalence of fish communities in our site.

4 | Discussion

Prevalence is a common parasitological parameter that gives information on the proportion of infected individuals in a given group (Bush et al., 1997; Rózsa et al., 2000). Our main goal here was to investigate the estimate of prevalence in fish communities across multiple spatial scales using data on the black spot disease. This approach allowed us to skim over spatial patterns of infection, sampling effort effect, sampling methods bias and, environmental predictors of the prevalence of infection in fish communities.

Despite that fish hosts have been the focus of many empirical studies on aquatic parasites, sampling design and methods bias are not usually discussed although they can induce substantial differences in measurements derived from wildlife samples (Biro & Dingemanse, 2009). For example, Wilson et al. (1993) reported difference in parasite infection in pumpkinseed sunfish (*Lepomis gibbosus*) caught by two different fishing methods. They mentioned that black spots were twice as abundant in trapped fish against seined fish, and white grubs twice as abundant in seined fish, indicating method-induced bias. Our findings showed prevalence estimate variations between sampling methods up to 16.4% at landscape-scale (Table S12) and 46.7% at lake-scale (Table S13). These results suggest substantial method-induced bias in the measurement of community-level infection prevalence. However, while transects measured the highest and minnow traps to lowest prevalence at landscape-scale, this tendency was not consistent at-lake-scale.

Sampling wildlife is challenging, and all methods come with their load of advantages and inconveniences. While fishing methods require a more extensive sampling than field observations (they sample fewer individuals), assessments of infection are accurate and precise because the observer can take the necessary time to inspect fish individuals. Minnow traps are typically selective on active, bold and curious behavior because fish must enter voluntarily in the trap (Biro

et al., 2006; Wilson et al., 1993). Besides, individual variations in behavior often lead to variation in infection risk (Hart, 1990), active and bold individuals usually associated to higher infection risk (Boyer et al., 2010; Santicchia et al., 2019; Wilson et al., 1993). Conversely, parasite infection can alter host behavior including habitat use and risk taking (Moore, 2002), making them more vulnerable to "trappability" (Barber & Dingemanse, 2010). In the end, it remains unclear if host individual behavior define their infection risk and/or whether parasite infection alters their personality (Dubois & Binning, 2022). Even if seine net is not directly a behavior-selective gear, this method comes with important habitat-selection restrictions. Sites must not be deeper than the height of the net to avoid fish escaping and, ideally, must not have large obstacles (e.g., branches, trunks, boulders) because the net would get stuck in it, resulting in tearing the net or, creating openings where fish can escape while researchers unstuck it. Consequently, seine sites might have similar habitat features despite effort to sample habitat heterogeneity. Since species have different habitat preferences, seine net might be indirectly selective on certain fish species and personality because of its habitat restrictions. For example, Biro et al., (2006) found that bold trout were more frequently caught in open risk habitat than shy trout, corroborating habitat use segregation based on intraspecific personality. However, habitat use can change in the context of competition and/or predation (Fischer, 2000; Mittelbach, 1986; Werner & Hall, 1977). Therefore, community composition must be considered when inferring species selection bias based on habitat use. Observational snorkeling transect is not a habitat-restricted nor a behavior-selective methodology. Although, quality of observations underwater relies on, meteorological conditions, water transparency (color, turbidity) and expertise of the observers to identify fish species in movement. As a result, precision of infection assessment can vary between sampling days and, between lakes. Low black spot abundance might be missed because of poor visibility or fish swift movement, leading to underestimation of prevalence estimates. Moreover, since infection assessment is easier in curious and solitary fishes as they usually get closer to the observers, accuracy of prevalence estimates might vary between fish species and personality. Finally, since older/bigger fish have a higher risk of infection than younger/smaller ones (Lemly & Esch, 1984).

Sampling effort must be sufficient to measure adequate infection metrics. However, parasitological parameters are often estimated with small sample size because of ethic, time, and monetary constraints. Our results showed that low sampling effort tend to overestimate the prevalence value at landscape-scale (Figure 6). This is because a high proportion of local sites had prevalence values above the regional average (see Table S14), thus at low sampling efforts the probability of randomly sampling a site with high prevalence is high. However, because many sites had a prevalence of zero, higher sampling efforts led to a quick decline in prevalence estimate until the true regional average was reached. Our results showed that low sampling effort lean toward overestimating the prevalence value at landscape-scale (Figure 6). If most samples inhere a prevalence value above the landscape prevalence estimate, there is a higher chance to sample those at low sampling effort thus overestimating the mean landscape prevalence. However, presence of few zero prevalence samples lower the prevalence estimated across the landscape as sampling effort increase. This evidence presence of high and low infection clusters (infection hotspots and coldspots). Parasites have an aggregated nature, where more individuals have low infection intensity, and few individuals are highly infected. This means highly infected individuals might occur in habitats favoring parasite occurrence and transmission. Thus, heterogenous landscape should present infection hotspots where prevalence is high as it correlates with parasite abundance (Šimková et al., 2002). The relationship between prevalence in fish communities and sampling effort have not been addressed before. Although, some researchers investigated the effect of sample size (number of sampled hosts) on prevalence estimate. Marques & Cabral (2007) found no significant effect of sample size on prevalence values, yet they reported that simulated median values tended to over or underestimate the real value at low sample size. Jovani & Tella (2006) also reported that low sample size resulted in greater inaccuracy and suggested an optimal sample size around 15 as a trade-off between losing data and maintaining accuracy. Our results evidence that minimum number of samples needed to adequately estimate landscape prevalence differed between methods. The only sampling method that reached a stable value was the observational snorkeling transects (around 10 samples) although the seine and methods combination showed less than 2% of difference between resampled prevalence and actual prevalence after 35 samples (Table S17). Indeed, snorkeling transects allow observations on more

individuals than the fishing methods, explaining why it reaches stability faster and thus, need fewer sampling effort to measure accurate prevalence. 35 sampling efforts were not exhaustive enough to capture the actual infection landscape prevalence (Table S17). Ultimately, «magic cutoff» sampling effort for estimating accurate prevalence does not exist as it can vary with parasite aggregation level (Poulin, 2007; Shvydka et al., 2018), host species target and abundance, sampling design (e.g., timing) and as we have shown, sampling method. However, our results should be used as an empirical tool to guide future determination of sampling effort to assess infection prevalence in natural systems.

Our results show that prevalence of infection is spatially heterogenous in our system (Figure 5). Prevalence estimates varied among and within lakes. Regardless, we did not find evidence of geographical patterns. Accordingly, none of the spatial attributes that we tested were good predictors of infection at small spatial scale (site-scale) (Table S18). Additional lakes would allow us to confirm whether infection is spatially random, or patterns could not be detected because of our small lake sample size (N=14).

Frequency distributions are a great tool to investigate mechanisms behind spatial occurrence of species (McGeoch & Gaston, 2002). For parasites, frequency occupation distributions have been revised so that host species represent suitable patches (i.e. habitat)(Hess et al., 2002). In our case, frequency distributions were made to investigate the distribution of prevalence of black spot disease infection in lake fish communities. Accordingly, histograms do no show parasite frequency of species but rather frequency of lakes' community with corresponding prevalence level. Results showed that frequency distributions differed between sampling methods, some suggesting a bimodal distribution tendency (Figure 5.A, B) and others with no obvious pattern (Figure 5.C, D). While binomial shapes are frequently observed in frequency occupation distributions, some argue that they are an artifact of small sampling scale (Nee et al., 1991). However, it revealed that lake community tend to display a high or low infection prevalence. Presuming that black spot trematodes dispersion is not limited by the final hosts' movements (birds) across our study area, these high- and low-prevalence groups could present environmental

characteristics that favor (or restrict) parasite abundance and transmission to fish hosts. Alternatively, this could mean that when the parasite is present, there are regulating pressures that limit the parasite abundance and transmission. Yet, if the parasites overcome the selective pressure, they become highly prevalent in the fish community. A larger lake sample size might clarify if these patterns of infection dynamics.

Differences in frequency distributions imply that method-biases are also perceptible at lake-scale and that they influence observed prevalence distribution patterns. Accordingly, maps (Figure 5) clearly show that lakes' prevalence can be over or underestimated according to the method. For example, in lake Pin rouge, the transect method underestimated the community prevalence while in lake Croche, seine net overestimated the community prevalence (see Table S13). This evidence method-biases in field sampling of fish communities. If we take back the previous example, this means that the proportion of infected fishes sampled by seine nets in lake Croche was higher than the proportion sampled by the other methods. The method might have selected fish species with high host specificity, sampled infection hotspots, sampled older individuals with higher infection exposure or sampled active and bold individuals that have a higher infection risk. Our approach revealed method-biases in prevalence estimates, but do not allow us to establish the mechanism behind this causality. Moreover, method-biases seem to be lake-dependent.

At the site-scale, we used GAMMs to inspect relationships between the prevalence estimate and environmental predictors (Figure 7, Table S18). Predictors are important to assess characteristics of infection hotspots and predict extinction risk with growing anthropogenic pressures. Our results evidence some non-linear patterns supporting the use of flexible approach to model empirical (data-driven) infection dynamics. All physico-chemistry parameters (turbidity, temperature, dissolved oxygen, conductivity, and pH) were good predictors of infection prevalence at site-scale (Figure 7.A, B, C, D, E). Comparatively to infection assessment of fish populations and communities by field surveys (or biological samples), water physico-chemistry measurements do not require a lot of equipment, time and are non-invasive to organisms. Thus, water physico-chemistry might be used as proxies of parasite infection for future monitoring and managing

strategies. Consequently, we encourage researchers to investigate these relationships in other parasite species, geographical localities, and types of habitats as our data only snapshot the complexity of fish-parasite associations in natural systems.

Surprisingly, habitat structure features were not good predictors of prevalence of infection in site-scale fish communities from the littoral zone. However, we found a strong (D^2 = 84.17%) negative relationship between the macrophyte cover and the prevalence of infection in transect sites (Figure 7.G). Contrary to our findings, heavier infection has been measured in fishes associated with vegetated zones in previous studies (Hartmann & Nümann, 1977; Marcogliese et al., 2001; Ondrackova et al., 2004). On one hand, because cercariae have a short life-span (Combes et al., 1994; Pietrock & Marcogliese, 2003), physical barriers might prevent the free-living stage from encountering a suitable host. Macroalgae have been shown to reduce the number of trematode cercariae (Bartoli & Boudouresque, 1997; Welsh et al., 2014) thus interfering with the parasite transmission to the next host. For instance, Prinz et al., (2009) demonstrated that inert seaweed can obstruct transmission of trematodes cercariae to their mussel host. Similarly, Buss et al., (2022) suggested that microplastic fibers might reduce success of infection by impeding cercariae to encounter their amphibian host. A treatment of complex shell surface of oysters also lead to reduction of 44% of cercariae in a experimental study (Welsh et al., 2014). In short, structural barriers might impede transmission by trapping cercariae, damaging them or consuming time of their short infective window. On the other hand, infected snails releasing cercariae in our lake system might not live or feed on macrophytes, explaining why we did not find a positive relationship between macrophyte cover and prevalence of infection in fishes. Indeed, if the association between macrophyte and snails don't stand, there is no reason to think that increasing macrophyte would lead to an increase in cercariae release and accordingly prevalence of infection in the fish community. For example, the mud Amnicola (*Amnicola limosus*) is the second intermediate host to at least one species of the black spot trematode guild in our lake system (Levet, unpublished). This freshwater snail species live partially in the lake sediment from the littoral zone (Pinel-Alloul & Magnin, 1973) and often graze on periphyton (Kesler, 1981). Consequently, we would not expect a strong relationship with macrophyte cover.

Our results suggest that community structure is a driving factor of prevalence of the black spot disease at small spatial scale. We found negative relationships with the total fish abundance, the non-host abundance and the Simpson's diversity index (Figure 7.J, K, L). These results support the dilution effect hypothesis. Since total fish abundance and non-host abundance are positively correlated (Figure S1), we assume similar mechanisms explaining why lower prevalence was measured in transect with higher fish abundance. First, a higher non-host fish abundance might reduce prevalence estimate in fish communities by an unsuccessful infection attempt on a non-host individual, by encountering individuals playing a structural barrier role (explained above) or by inducing a behavior change that reduces encounter rate with fish hosts. For example, Ahn & Goater (2021) showed that abundance of brain worms decreased in fathead minnows (*Pimephales promelas*) when mixed with non-host emerald shinners (*Notropis atherinoides*), but that there was no difference when mixed with other non-host species, suggesting a behavioral change when *P. promelas* and *N. atherinoides* co-occur. This emphasis fish species identity role in dilution effect mechanisms. Accordingly, diverse communities in our system tend to be less infected than communities with dominant species. In our lake system, *L. gibbosus* is a dominant species of littoral communities and the species with the higher infection prevalence across the landscape (Table S19). The presence of competitor species (e.g., *P. flavescens*) might then decrease the overall prevalence in the community by modifying pumpkinseeds sunfish's behavior that increases encounter with cercariae. Indeed, Dargent et al. (2013) found that presence and abundance of *Gyrodactylus* spp. (monogenean) were lower in hosts when they occur in mixed-species groups. Even though it was not considered in this study, local diversity of non-fish organisms could influence the infection in fish. For instance, Lagrue & Poulin, (2015) found a negative association of non-host benthic invertebrates diversity with infection prevalence in second intermediate host. In fact, non-host species can reduce infection success of cercariae by actively or passively feeding on them (Thieltges et al., 2013). This has been observed in many taxa in aquatic ecosystems including zooplankton, Bivalvia, crustacean macroinvertebrates, cnidarian, and small fish (Kaplan et al., 2009; E. Mironova et al., 2019; K. Mironova et al., 2020; Schotthoefer et al., 2007; Vielma et al., 2019).

Understanding the process explaining dynamics of parasitic disease across scale is fundamental to understand future infection and extinction risk in natural systems. Here we investigated prevalence of the black spot disease across multiple scales to skim over sampling effort relationship, sampling method bias, spatial occurrence patterns and environmental predictors. Results suggested that prevalence is distributed in a non-random heterogeneous way in the landscape with small sampling size largely overestimating regional prevalence. We evidenced inconsistent method-induced bias in prevalence estimates at both lake and landscape-scale, sometimes leading to high variation in estimates. Method also influenced sampling effort needed to reach accurate landscape prevalence estimate, the observational snorkeling transects requiring the least. Best predictors of site-scale prevalence were water physico-chemistry and community structure metrics. Our findings suggest encounter dilution effects by host decoy or physical obstruction. However, we observed many non-linearities in the effects suggesting complex interactions that we could not disentangle with our approach. Therefore, we encourage research to deepen knowledge to understand processes linking these predictors to infection parameters for future integration into sampling surveys and conservation plans. In short, we evidenced variations of infection prevalence across spatial scale, sampling effort and sampling methods. Hence, we urge researchers to take these into consideration in the context of comparative studies.